WHERE DOES A REPRESENTATION'S VALUE LIVE? THE REALIZABILITY DECOMPOSITION OF DIMENSIONALITY REDUCTION AND DEPTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Methods grouped under *dimensionality reduction* are used to obtain lower-dimensional descriptions of data, but "fewer coordinates" does not specify the geometric operation that produced them. Against the reference of a basis change, which keeps all D coordinates, we separate two constructions that both return $k < D$ numbers: a projection onto a subspace of the *original* ambient space, and an embedding into a *distinct* target space. Every rank-k linear map is of the first kind, an orthogonal projection followed by an invertible coordinate change with a closed-form metric that restores the projected geometry, so the distinction has content only for nonlinear and learned maps. To measure it we introduce the *realizability decomposition*: any finite centered representation Z of a declared source X splits orthogonally as $Z = Z_{\|} + Z_{\perp}$, where $Z_{\|}$ is exactly the part expressible as coordinates of one fixed source-subspace projection and $Z_{\perp}$ is what no such projection can produce. The residual $\rho = \|Z_{\perp}\| / \|Z\|$ comes with an exact finite-sample caveat and a random-representation null. The decomposition also makes precise what a linear autoencoder identifies (a subspace and a metric, never a basis) and why a decoder trained with a pointwise loss returns to the original space rather than a symmetric copy of it. Empirically, a fixed projection of pixel space carries 82–95% of the energy of t-SNE, UMAP and PaCMAP layouts of MNIST and Fashion-MNIST, and 91–95% on CIFAR-10, yet that part has only PCA-level neighborhood fidelity; the embeddings' advantage lives in the small remainder $Z_{\perp}$. For autoencoders, end-to-end training produces codes with $Z_{\perp} \neq 0$ whenever it beats PCA, but a nonlinear decoder trained on PCA scores ($Z_{\perp} = 0$) beats PCA as well, so the remainder describes the code training chose, not a requirement of reconstruction. The decomposition is a diagnostic, not a theory of representation: it says how much of a representation is a projection of a declared source, and what follows from that.

1 INTRODUCTION

Take a t-SNE layout of MNIST (van der Maaten & Hinton, 2008). Center the pixels and the layout, and regress the layout's two coordinates on the 784 pixel values. The fit captures 83% of the layout's energy (its squared Frobenius norm, so this is the regression's R^2), and the fitted part is, by a theorem proved below, exactly a fixed orthogonal projection of pixel space read through a linear change of coordinates. Evaluated as a representation, that fitted part has trustworthiness 0.77, the level of PCA (0.75), and a 10-nearest-neighbor classifier on it reaches 58%, above PCA's 44% because the projection it realizes is not PCA's, but far below the full layout's trustworthiness 0.98 and accuracy 94%. What separates a projection-grade representation from a t-SNE-grade one lies in the remaining 17% of the layout's energy, which no projection of the source can produce.

This paper is about that split. PCA, SNE, t-SNE, UMAP, Isomap, LLE and autoencoders (Jolliffe & Cadima, 2016; Hinton & Roweis, 2002; McInnes et al., 2018; Tenenbaum et al., 2000; Roweis & Saul, 2000; Hinton & Salakhutdinov, 2006) are grouped under one phrase because each replaces D coordinates by $k < D$. The inequality does not say what happened to the data: a projection moves each point inside the same ambient space and reports its coordinates, while an embedding places each sample in a different space by an objective, with nothing tying it to a fixed view of the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

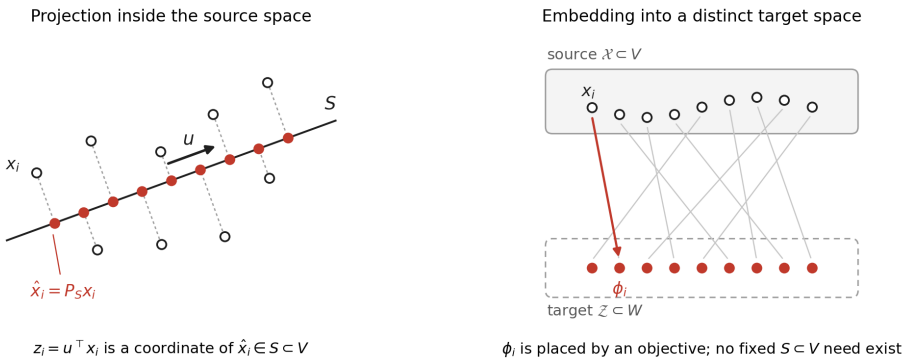

Figure 1: Two operations that both produce fewer coordinates. **Left:** orthogonal projection moves each point to $\hat{x}_i = P_S x_i$, which remains in the source space; the coordinate of $\hat{x}_i$ along u is a description of that point. **Right:** a cross-space embedding assigns each sample a target-space location chosen by an objective; nothing constrains it to be a coordinate of a fixed source-space projection.

source (Figure 1). Recent work treats learned representations as geometric objects, asking what their dimension is (Chun et al., 2026), what manifolds they form (Moisescu-Pareja et al., 2026), what geometry an objective induces (Alshammari et al., 2025; Gupta et al., 2024), and how much a map distorts distances (Uscidda et al., 2025). We ask a complementary question: of a given representation, *how much is a projection of its source, and is that the part that matters?*

Contributions. (1) Every rank-k linear reduction is a same-space projection followed by an invertible coordinate change, with a metric under which the reported coordinates are isometric to the projected geometry (§3). (2) The realizability decomposition $Z = Z_{\|} + Z_{\perp}$: $Z_{\|}$ is the best fixed-projection approximation of Z, ρ is a least-squares residual with an exact finite-sample failure and a null, and the construction applies to any declared source (§4). (3) What an autoencoder's latent space identifies, why its decoder returns to the original space and points, and a local structural check (§5). (4) On MNIST, Fashion-MNIST and CIFAR-10, the neighborhood fidelity of t-SNE, UMAP and PaCMAP lives in $Z_{\perp}$; autoencoder gains over PCA accompany $Z_{\perp}$ in end-to-end training but do not require it (§6). (5) Local and layer-wise versions of the decomposition, and a certificate for replacing a network layer by an affine map (§7, Appendices I–J).

2 Three Operations Behind "Dimensionality Reduction"

Let $V = \mathbb{R}^D$, $k < D$, and $X \subset V$ a data set. A **basis change** $z = U^\top x$ with $U \in \mathbb{R}^{D \times D}$ orthogonal keeps the vector and all D coordinates; it is the reference. A **projection within the original space** maps x to $\hat{x} = P_S x$ with $P_S = U_k U_k^\top$ for a k-dimensional subspace S with orthonormal basis U_k: the vector changes but the ambient space does not, and the coordinates $z = U_k^\top x = U_k^\top \hat{x} \in \mathbb{R}^k$ are a code for the source-space point $\hat{x} = U_k z$. A **cross-space embedding** $f : \mathcal{X} \subset V \to \mathcal{Z} \subset W$ places points in a target space by an objective; SNE preserves neighborhood probabilities (Hinton & Roweis, 2002), and I-Con treats the SNE mapper as a lookup table (Alshammari et al., 2025). Nothing in the construction identifies a subspace $S \subset V$ whose coordinates it reports. Table 1 lists common maps under both attributes, and Appendix B works a two-dimensional example of all three operations. The question this paper answers is how to tell, for a learned f, which case one is in, and how much of f is each.

3 Every Linear Reduction Is a Projection

Theorem 1 (Coordinate isometry). *If U_k has orthonormal columns and $S = \mathrm{span}(U_k)$, the coordinate map $c(s) = U_k^\top s$ is an isometry from S with the inherited Euclidean metric onto $\mathbb{R}^k$.*

Theorem 2 (Linear reduction factorization). *Let $W \in \mathbb{R}^{k \times D}$ have rank k. There exist $U \in \mathbb{R}^{D \times k}$ with orthonormal columns and an invertible $A \in \mathbb{R}^{k \times k}$ with $W = AU^\top$. Hence $z = Wx$ is an invertible coordinate transformation of the orthogonal projection of x onto $S = \mathrm{row}(W)$: with*

Table 1: Taxonomy. "Coords. ↓" asks whether the number of reported coordinates decreases; the geometric interpretation is the separate attribute this paper measures. $P_S x$ is still a point of $\mathbb{R}^D$ (D coordinates, rank k); only a coordinate chart of S, as in PCA scores, reports k numbers.

Operation	Map	Coords. ↓	Geometric interpretation	Preserved structure
Basis change	$x \mapsto U^\top x$, U orthogonal	No	Same vector, new coordinates	Exact
Subspace projection	$x \mapsto P_S x$	No (D, rank k)	Same ambient space, in S	Projected geometry
PCA scores	$x \mapsto U_k^\top x$	Yes	Coordinates of $P_S x$	Isometric to projection
General linear encoder	$x \mapsto Wx$	Usually	Projection onto $\mathrm{row}(W)$, oblique coords	Metric $G = A^{-\top}A^{-1}$
SNE / t-SNE / UMAP	$x_i \mapsto \phi_i$	Yes	Target embedding space	Neighborhood distributions
Nonlinear autoencoder	$x \mapsto E(x)$	Usually	Learned latent space	Reconstruction + inductive bias
Kernel feature map	$x \mapsto \phi(x)$	Often ↑	Target feature space	Kernel inner products

$\hat{x} = UU^\top x$, $z = AU^\top \hat{x}$ and $\hat{x} = UA^{-1}z$. Constructively, $U = Q$ and $A = R^\top$ from the thin QR factorization $W^\top = QR$.

Corollary 1. *W is itself an orthonormal coordinate map of a projection if and only if $WW^\top = I_k$, in which case $P_S = W^\top W$.*

Proofs of all results are in Appendix A. The content of Theorem 2 is what it rules out: no linear map can *produce* a representation that is not the coordinates of a same-space projection, whatever objective motivated it. PCA is the case $A = I$; a random projection, a linear autoencoder's encoder and a linear read-out differ from it only in the coordinate system A. That coordinate system distorts Euclidean geometry: with orthonormal subspace coordinates $c_i = U^\top x_i$,

$$\|z_i - z_j\|^2 = (c_i - c_j)^\top A^\top A\,(c_i - c_j), \qquad \frac{\|z_i - z_j\|}{\|\hat{x}_i - \hat{x}_j\|} \in [\sigma_{\min}(A), \sigma_{\max}(A)], \tag{1}$$

but the distortion is a property of the coordinates, not of the projection, and it is undone by the metric

$$G = A^{-\top}A^{-1}: \qquad (z_i - z_j)^\top G\,(z_i - z_j) = \|P_S x_i - P_S x_j\|^2. \tag{2}$$

so a linear encoder with $\kappa(A) \gg 1$ has large Euclidean distortion and none under the right metric, the exact linear instance of the distortion question of Uscidda et al. (2025). Since every linear map is a projection, a quantity that separates projections from embeddings must measure *how much* of a nonlinear map is a projection.

4 The Realizability Decomposition

Definition 1 (Projection-realizable representation)**.** $f : \mathcal{X} \subset \mathbb{R}^D \to \mathbb{R}^k$ is projection-realizable on $\mathcal{X}$ within a class $\mathcal{C}$ of coordinate maps if there exist a k-dimensional $S \subset \mathbb{R}^D$ and a bijection $c : S \to \mathbb{R}^k$ in $\mathcal{C}$ with $f(x) = c(P_S x)$ for all $x \in \mathcal{X}$. For $\mathcal{C}$ linear, similarity (rotation and scale), or orthonormal, we say linearly, similarity, or orthogonally realizable.

The class must be declared: on a finite set, unrestricted c manufactures a factorization for any f. Let $X \in \mathbb{R}^{n\times D}$ and $Z \in \mathbb{R}^{n\times k}$ be centered, and let $\Pi_X = XX^+$ be the orthogonal projector in $\mathbb{R}^n$ onto $\mathrm{col}(X)$.

Proposition 1 (Realizability decomposition)**.** *Every centered Z splits uniquely and orthogonally as*

$$Z = Z_\parallel + Z_\perp, \qquad Z_\parallel = \Pi_X Z = XW, \quad W = X^+ Z, \qquad Z_\perp \perp \mathrm{col}(X), \tag{3}$$

where (i) the fitted matrix $Z_\parallel$ is unique: $\min_M \|Z - XM\|_F$ is attained exactly on the affine set $X^+Z + \{N : XN = 0\}$, every element of which gives $XM = Z_\parallel$, so the coefficient matrix is unique iff $\ker X = 0$ and $W = X^+Z$ is its minimum-norm choice; and $\inf\{\|Z - XUA^\top\|_F : U^\top U = I_k,\ A \text{ invertible}\} = \|Z_\perp\|_F$, the infimum being attained in the constrained class if and only if $\mathrm{rank}(X^+Z) + \dim\ker X \geq k$ (always when $\mathrm{rank}\,X \leq D - k$, as for MNIST with $\mathrm{rank}\,X = 642$; exactly when $\mathrm{rank}(X^+Z) = k$ if X has full column rank); (ii) the data-identified source subspace is $S_{\text{eff}} = \mathrm{span}(X^+Z)$, of dimension $\mathrm{rank}(X^+Z) \leq k$, with orthonormal basis $U = \mathrm{orth}(X^+Z)$; when $\mathrm{rank}(X^+Z) = k$ it is the unique k-dimensional realized subspace, and when the rank is smaller a rank-k minimizer can only add directions in $\ker X$, which leave the fitted representation on

the sample unchanged and are not identifiable from the data; (iii) no linear function of the source has any component along $Z_\perp$; and (iv) the residual

$$\rho_{\text{lin}}(X, Z) = \frac{\|Z_\perp\|_F}{\|Z\|_F} \in [0, 1], \qquad \rho^2_{\text{lin}} = 1 - R^2, \tag{4}$$

where R^2 is the total coefficient of determination of the multivariate least-squares regression of Z on X, satisfies $\rho_{\text{lin}} = 0$ iff Z is linearly projection-realizable.

ρ_{lin} is the residual of a multivariate linear regression, the same quantity as the linear-regression index of Kornblith et al. (2019). What the decomposition adds is the reading and the objects: by Theorem 2 the fitted part $Z_\parallel$ is exactly the coordinates of one orthogonal projection of the source, so ρ_{lin} is the fraction of a representation that is not a projection, and $Z_\parallel$ and $Z_\perp$ can each be evaluated as representations, which a similarity score cannot. Appendix C compares ρ with CKA, CCA and orthogonal-Procrustes shape distances (Raghu et al., 2017; Williams et al., 2021): none of them separates exact projections from embeddings (CKA gives 0.68 to PCA and 0.20 to a random orthonormal projection, both exactly realizable), and full-rank CCA gives 0.36 to pure noise.

Proposition 2 (A linear decoder cannot use $Z_\perp$)**.** *For centered X and Z, let $e(Z) = \min_W \|X - ZW\|_F^2$ be the error of the best linear decoder from Z. Then*

$$e(Z) \;\geq\; e(Z_\parallel) \;\geq\; e_{\text{PCA-}k}(X), \tag{5}$$

where $e_{\text{PCA-}k}$ is the rank-k PCA reconstruction error. Consequently an autoencoder with any encoder and a linear *decoder cannot beat PCA at the same k under squared loss, and its non-realizable component can only hurt. The statement is about linear decoders only: a nonlinear decoder can beat PCA from a code with $Z_\perp = 0$ (for points on a parabola $x_2 = x_1^2$, the projection $z = x_1$ is exactly realizable and decodes without error), and §6.2 shows that this happens on image data too. Whether a gain over PCA* uses *$Z_\perp$ is therefore an empirical question about a trained autoencoder, not a consequence of this proposition.*

We also use the similarity residual $\rho_{\text{sim}} = \min_{Q^\top Q = I_k,\, s>0} \|Z - sXQ\|_F \,/\, \|Z\|_F \geq \rho_{\text{lin}}$, which asks for realizability up to rotation and units, i.e. up to Euclidean geometric equivalence; it is a nonconvex Stiefel problem solved by projected gradient descent from the orthonormalized W (Appendix M). We do not use the residual with $s = 1$: t-SNE and UMAP report coordinates in arbitrary units, so a fixed scale would measure units, not geometry.

Proposition 3 (Finite-sample non-identifiability)**.** *If centered X has rank $n - 1$, which holds generically when $n \leq D + 1$, then $\rho_{\text{lin}}(X, Z) = 0$ for every centered Z.*

With fewer samples than ambient dimensions, every representation, including t-SNE of pure noise, is linearly realizable. This is the same phenomenon that biases finite-sample dimension estimates (Chun et al., 2026); below $n \approx D$ the sample cannot distinguish a structured map from an unstructured one. We therefore always report ρ_{lin} with a *random-representation null*: ρ_{lin} of a Z with i.i.d. standard Gaussian entries drawn independently of X, with the same n, D, k, for which $\mathbb{E}\,\rho^2_{\text{lin}} = 1 - \text{rank}(X)/(n-1)$ exactly (Appendix A), i.e. 0.84–0.87, so that the null residual itself is $\rho_{\text{lin}} \approx 0.92$–$0.93$ on MNIST and Fashion-MNIST (0.83 on CIFAR-10).

Relative realizability. Nothing in (3) requires X to be raw data. Replacing X by a feature map $\phi(X)$, or by the activations $H_{\ell-1}$ of the previous layer of a network, gives $\rho(\phi(X), Z)$ and $\rho(H_{\ell-1}, H_\ell)$: how much of Z is a fixed projection of *that* source. When ϕ contains the identity features the residuals are nested, $\rho(\phi(X), Z) \leq \rho(X, Z)$, and the difference is the geometry only the richer source accounts for.

5 What an Autoencoder Identifies, and Where It Returns To

Theorem 3 (Latent reparameterization invariance)**.** *Let $E : \mathcal{X} \to \mathcal{Z}$, $\mathcal{D} : \mathcal{Z} \to \mathcal{X}$, and let $h : \mathcal{Z} \to \mathcal{Z}'$ be any invertible map. With $E' = h \circ E$ and $\mathcal{D}' = \mathcal{D} \circ h^{-1}$, $\mathcal{D}'(E'(x)) = \mathcal{D}(E(x))$ for all x, while latent distances and angles may change arbitrarily.*

For a *linear* autoencoder, Theorem 2 writes the encoder as $W_e = AU^\top$: what is determined is the subspace $S = \text{span}(U)$ and the metric G of (2), and the latent axes only up to A. With squared

loss, S at the global optimum is the principal subspace (Baldi & Hornik, 1989; Plaut, 2018), so the code is PCA scores read through A. The subspace projected *onto* is $\mathrm{col}(W_d)$, recovered from the decoder; the encoder's row space coincides with it only when $W_d W_e$ is an orthogonal projector, and the trained ones here are oblique (Appendix D). For a *nonlinear* autoencoder there is no basis to know: the latent space is not a subspace of the input space, the local frames given by the encoder and decoder Jacobians move with x, and Theorem 3 leaves latent axes, distances and angles unidentified unless a symmetry-breaking term is added. What can be said is quantitative: $Z_{\|}$ is the part of the code that is a fixed projection of the input, with basis $\mathrm{orth}(X^{+}Z)$, and $Z_{\perp}$ has no basis in the input space by construction.

Proposition 4 (Output-space identification)**.** *Let $\mathcal{L}(\theta) = \sum_i d\big(x_i, \mathcal{D}_\theta(E_\theta(x_i))\big)$ with $d(x, y) = 0$ iff $x = y$. (i) Every latent reparameterization leaves $\mathcal{L}$ and $\mathcal{D} \circ E$ unchanged. (ii) If $\mathcal{D} \circ E$ is exact on the data, then $g \circ \mathcal{D} \circ E$ has strictly larger loss for every isometry g of $\mathbb{R}^D$ that moves some data point. If instead $\mathcal{L}$ sees the outputs only through a statistic invariant under a group $\mathcal{G}$, then $g \circ \mathcal{D} \circ E$ has the same loss for every $g \in \mathcal{G}$.*

The output space is fixed by construction (the decoder emits D numbers) and its coordinate labels by the loss, which compares those numbers entry by entry with x; the latent symmetry cancels before the output. A sharper worry is a copy of the data that carries the *same values and the same representation*: a reflection, an involution, or any bijection φ that preserves every pairwise distance and neighborhood, which a neighborhood or distribution objective cannot see. A pointwise objective can.

Proposition 5 (Uniqueness of the identity reconstruction)**.** *With d as above, $R = \mathcal{D} \circ E$ minimizes $\mathbb{E}_P\, d(x, R(x))$ iff $R(x) = x$ for P-a.e. x (on a finite sample: on the training set). Hence every bijection φ with $\varphi \neq \mathrm{id}$ on a set of positive measure (a non-trivial isometry, an involution, a map preserving all distances and neighborhoods) has strictly positive loss as a reconstruction, even when P is φ-invariant.*

The decoder is thus a left inverse of the encoder on the image of the data only; between training samples R is fixed by inductive bias, so the operational check is the *held-out* reconstruction error. A local, structural check (full statements and proofs of Propositions 4–6 in Appendix A) distinguishes a projection onto the decoded manifold from a symmetric copy of it.

Proposition 6 (Projector structure of the reconstruction map)**.** *Let $R = \mathcal{D} \circ E$ be C^1, with decoded manifold $M_{\mathcal{D}} = \mathcal{D}(\mathcal{Z})$. (i) Latent cycle-consistency, $E \circ \mathcal{D} = \mathrm{id}$ on $E(\mathbb{R}^D)$, implies $R \circ R = R$, and conversely when $\mathcal{D}$ is injective on $E(\mathbb{R}^D)$. (ii) If $R \circ R = R$, then at every $x^\star = R(x)$ the Jacobian $J = dR_{x^\star}$ is idempotent with spectrum in $\{0, 1\}$; if also $\mathcal{D}$ is an immersion at $E(x^\star)$ and $\mathrm{rank}\, dE_{x^\star} = k$, then $\mathrm{rank}\, J = \mathrm{tr}\, J = k$ and $\mathrm{col}(J) = T_{x^\star} M_{\mathcal{D}}$, so R is locally an (oblique) projection onto $M_{\mathcal{D}}$. (iii) If instead $R = g \circ R_0$ with R_0 as in (ii) and g a non-trivial isometry with $g(M_{\mathcal{D}}) = M_{\mathcal{D}}$, then at a point $x^\star = R_0(x^\star)$ fixed by g, $dR_{x^\star}$ has an eigenvalue of modulus one other than 1 unless dg is the identity on $T_{x^\star} M_{\mathcal{D}}$; away from fixed points of g no such eigenvalue need exist.*

For a nonlinear autoencoder we evaluate dR analytically at data points and report its idempotence defect, the number of eigenvalues near 1, the number of eigenvalues of modulus $> \frac{1}{2}$ with negative real part (a necessary check, by part (iii), not a certificate), and the map-level defect $\|R(R(x)) - R(x)\| \,/\, \|R(x)\|$. Reconstruction loss alone does not enforce part (i): a decoder can reconstruct the training set exactly while $E \circ \mathcal{D} \neq \mathrm{id}$ off the latent image. Finally, the last clause of Proposition 4 is why the coordinate class of Definition 1 must be declared per objective: distribution-level objectives (adversarial, Wasserstein, Gromov–Monge, or SNE's KL) are invariant to every transformation preserving the compared statistic, so for SNE-type objectives ρ_{sim}, not ρ_{lin}, respects the objective's own invariance (Alshammari et al., 2025).

6 Where the Value Lives

All experiments use fixed seeds and released scripts (Appendix N). Real-data runs use $n = 5000$ test images of MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017) ($D = 784$, $n/D \approx 6$, so Proposition 3 does not apply), and the 10,000 CIFAR-10 test images (Krizhevsky, 2009) ($D = 3072$). The random-representation null (Gaussian Z drawn independently of X) is reported alongside every residual. On synthetic data (Appendix E) the linear theory holds to machine

Table 2: Realizability decomposition on $n = 5000$ test images ($D = 784$). $\rho = \rho_{\text{lin}}$. Trust. and kNN are reported for the full representation Z, its realizable part $Z_{\parallel}$, and its remainder $Z_{\perp}$ (— when $Z_{\perp} \approx 0$). θ: mean principal angle (degrees) between the realized subspace $\text{span}(X^{+}Z)$ and the PCA-k subspace, measured in score space (Appendix N).

Data	Method	k	ρ	Trust. Z	Trust. $Z_{\parallel}$	Trust. $Z_{\perp}$	kNN Z	kNN $Z_{\parallel}$	kNN $Z_{\perp}$	θ
MNIST	PCA	2	$< 10^{-14}$	0.747	0.747	—	0.437	0.437	—	0°
	Random orthonormal proj.	2	$< 10^{-14}$	0.580	0.580	—	0.193	0.193	—	73°
	Linear autoencoder	2	$< 10^{-14}$	0.747	0.747	—	0.434	0.434	—	2°
	Nonlinear autoencoder	2	0.011	0.748	0.747	0.518	0.441	0.441	0.150	1°
	t-SNE	2	0.409	0.981	0.767	0.511	0.939	0.575	0.218	32°
	UMAP	2	0.420	0.961	0.755	0.520	0.931	0.622	0.232	36°
	PaCMAP	2	0.427	0.942	0.745	0.513	0.932	0.573	0.261	32°
	null (random Z)	2	0.934	0.500	0.510	0.500	0.115	0.135	0.108	—
	PCA	10	$< 10^{-14}$	0.989	0.989	—	0.889	0.889	—	0°
	Random orthonormal proj.	10	$< 10^{-14}$	0.854	0.854	—	0.619	0.619	—	55°
	Linear autoencoder	10	$< 10^{-14}$	0.990	0.990	—	0.898	0.898	—	10°
	Nonlinear autoencoder	10	0.207	0.995	0.992	0.587	0.923	0.909	0.315	18°
	UMAP	10	0.415	0.985	0.912	0.587	0.939	0.817	0.427	41°
	null (random Z)	10	0.934	0.499	0.585	0.490	0.113	0.187	0.099	—
Fashion	PCA	2	$< 10^{-14}$	0.914	0.914	—	0.523	0.523	—	0°
	Random orthonormal proj.	2	$< 10^{-14}$	0.609	0.609	—	0.238	0.238	—	63°
	Linear autoencoder	2	$< 10^{-14}$	0.915	0.915	—	0.515	0.515	—	1°
	Nonlinear autoencoder	2	0.157	0.944	0.917	0.522	0.631	0.553	0.155	19°
	t-SNE	2	0.227	0.988	0.918	0.513	0.777	0.627	0.151	26°
	UMAP	2	0.227	0.979	0.903	0.528	0.753	0.641	0.163	35°
	PaCMAP	2	0.243	0.974	0.894	0.525	0.725	0.615	0.169	31°
	null (random Z)	2	0.918	0.502	0.514	0.503	0.087	0.123	0.101	—
	PCA	10	$< 10^{-14}$	0.993	0.993	—	0.773	0.773	—	0°
	Random orthonormal proj.	10	$< 10^{-14}$	0.936	0.936	—	0.631	0.631	—	48°
	Linear autoencoder	10	$< 10^{-14}$	0.993	0.993	—	0.779	0.779	—	1°
	Nonlinear autoencoder	10	0.124	0.995	0.994	0.578	0.804	0.790	0.260	16°
	UMAP	10	0.240	0.987	0.960	0.616	0.773	0.707	0.294	36°
	null (random Z)	10	0.918	0.498	0.556	0.492	0.111	0.187	0.113	—

precision, and the true coordinates of a Swiss roll have $\rho_{\text{lin}} = 0.40$: the right answer for a curved manifold is itself not globally realizable, which motivates the local version of Appendix I.

6.1 Neighborhood embeddings: most of the energy, little of the value

For each method we compute Z, its realizable part $Z_{\parallel} = \Pi_X Z$ and its remainder $Z_{\perp}$, and evaluate all three by trustworthiness at 10 neighbors (Venna & Kaski, 2001) and held-out 10-NN accuracy (Table 2). Methods: PCA, a random orthonormal projection, linear and nonlinear ($784-256-k-256-784$, $\tanh$) autoencoders, t-SNE, UMAP, PaCMAP (Wang et al., 2021), and the null.

Linear methods are exactly realizable ($\rho_{\text{lin}} < 10^{-14}$), as Theorem 2 requires; the random projection shares PCA's residual and none of its quality, so ρ measures *whether* a representation is a projection, not whether it is a good one. The linear autoencoder is PCA in oblique coordinates (Appendix D). **Neighborhood embeddings are mostly projection by energy, not by neighborhood fidelity.** t-SNE, UMAP and PaCMAP have $\rho = 0.41$–0.43 on MNIST and 0.23–0.24 on Fashion-MNIST, far below the nulls (0.93, 0.92), so their realizable parts carry 82–83% and 94–95% of the energy. That part matches PCA in trustworthiness (0.75–0.77 against 0.75 on MNIST, 0.89–0.92 against 0.91 on Fashion-MNIST) while the full layouts reach 0.94–0.99. In class separability it sits between the two (10-NN 0.57–0.62 against PCA's 0.44 and the layouts' 0.93–0.94 on MNIST), because the projection it realizes is not PCA's (mean principal angle $26°$–$36°$). At $k = 2$ the remainder alone is near chance; neighborhood fidelity is in neither part alone, and it is lost the moment $Z_{\perp}$ is removed. PaCMAP, which adds mid-near pairs to keep global structure, splits exactly like t-SNE and UMAP. Seed variation is below 0.01 in ρ for the embeddings (Appendix F).

The split is relative to the source, and part of it survives richer sources. Replacing pixels by random ReLU or RBF Nyström features and evaluating the realizable part out of sample (Appendix F), a linear expansion of the pixels to 2000 features changes nothing, while RBF Nyström features at matched dimension ($p = 784$) already lower ρ_{out} for t-SNE on MNIST from 0.47 to 0.31 and raise the realizable part's held-out accuracy from 0.52 to 0.70. About half of the gap between $Z_{\parallel}$ and Z is

Table 3: Realizability decomposition on CIFAR-10 ($n = 10{,}000$ test images, $D = 3072$, pixels in $[0, 1]$), as in Table 2. Trustworthiness is evaluated on a fixed random subset of 5000 images. The null residual is $\sqrt{1 - 3072/9999} = 0.83$.

Method	k	ρ	Trust. Z	Trust. $Z_\parallel$	Trust. $Z_\perp$	kNN Z	kNN $Z_\parallel$	kNN $Z_\perp$	θ
PCA	2	$< 10^{-14}$	0.819	0.819	—	0.173	0.173	—	0°
Random orthonormal proj.	2	$< 10^{-14}$	0.572	0.572	—	0.130	0.130	—	74°
Nonlinear autoencoder	2	0.346	0.853	0.812	0.515	0.187	0.161	0.106	13°
t-SNE	2	0.304	0.903	0.804	0.521	0.294	0.165	0.104	18°
UMAP	2	0.228	0.837	0.811	0.527	0.215	0.172	0.116	13°
PaCMAP	2	0.252	0.831	0.808	0.523	0.210	0.174	0.108	15°
null (random Z)	2	0.834	0.501	0.501	0.500	0.090	0.111	0.103	—
PCA	10	$< 10^{-14}$	0.977	0.977	—	0.300	0.300	—	0°
Random orthonormal proj.	10	$< 10^{-14}$	0.874	0.874	—	0.238	0.238	—	52°
Nonlinear autoencoder	10	0.068	0.973	0.974	0.536	0.313	0.318	0.117	4°
UMAP	10	0.226	0.908	0.901	0.561	0.266	0.237	0.120	46°
null (random Z)	10	0.832	0.502	0.530	0.483	0.099	0.101	0.105	—

nonlinear structure a generic kernel of pixel space contains; the other half is what the layout objective adds.

Natural images. On CIFAR-10 (Table 3) the layouts are *more* realizable than on digits ($\rho = 0.23$–0.30, so 91–95% of their energy), yet the realizable part is PCA-grade in both utilities: trustworthiness 0.80–0.81 against PCA's 0.82 and 10-NN 0.165–0.174 against 0.173, with realized subspaces $13°$–$18°$ from PCA's. Everything t-SNE adds (trustworthiness 0.90, 10-NN 0.29) is carried by $Z_\perp$.

6.2 Autoencoders: the gain over PCA accompanies $Z_\perp$ but does not need it

A sweep of 24 nonlinear autoencoders (depth 1–3, $\tanh$ and ReLU, $k \in \{2, 10, 32\}$; Appendix G, Figure 2) shows the co-occurrence: the only configuration with $\rho \approx 0$ (MNIST, $k = 2$, one $\tanh$ layer, $\rho = 0.015$) matches PCA's held-out error exactly, every configuration with $\rho > 0.05$ beats PCA, and for $\tanh$ at $k \geq 10$ the gain grows with ρ. At $k = 2$, and for ReLU, the two decouple. Proposition 2 does not make $Z_\perp$ necessary for a nonlinear decoder, and Table 4 tests whether it is. A nonlinear decoder of the same architecture trained on the PCA scores, a code with $Z_\perp = 0$ by construction, beats PCA in all six settings (gain 0.023–0.095) and matches or exceeds the end-to-end autoencoder at $k = 10$ and 32. End-to-end training helps at the tightest bottleneck, where its own code gains more than the PCA scores (0.065 vs. 0.025 on MNIST and 0.094 vs. 0.064 on Fashion-MNIST at $k = 2$), and the code it chooses has $Z_\perp \neq 0$. Its decoder then needs both parts: fed only $Z_\parallel$ it loses much of its gain (and falls below PCA on MNIST at $k = 2$ and 32), and fed only $Z_\perp$, placed at the code's mean, it does worse than predicting the mean image (relative error 1.01–1.25). The remainder is not a separate channel of information but a correction that is useful only on top of the projection. So ρ measures how far a trained code departs from a projection, and the gain over PCA is a joint property of the decoder's nonlinearity and the code it is given; $Z_\perp \neq 0$ is neither necessary nor sufficient. Latent regularization moves the code toward realizability: a strong contractive penalty (Rifai et al., 2011) drives ρ at $k = 10$ from 0.16 to 0.007 (MNIST) and from 0.10 to 0.008 (Fashion-MNIST) and removes the gain with it, while a β-VAE (Higgins et al., 2017) with $\beta = 4$ lowers ρ on MNIST and falls below PCA at $k = 10$ (Appendix H).

The recovery checks of Proposition 6 separate two guarantees that are easy to conflate. No autoencoder shows a mirror image: the count of eigenvalues with modulus $> \frac{1}{2}$ and negative real part is 0 at $k \geq 10$ and at most 0.07 per point at $k = 2$, and held-out errors below PCA's rule out a reconstruction $g \circ R$ with g a non-trivial isometry, which would pay $\|x - g(x)\|$ at every point it moves (Proposition 5). But recovering every point is not recovering the space. One-layer autoencoders are close to projections onto their decoded manifold (idempotence defect 0.07–0.38, Jacobian trace 1.9, 9.3, 31.9 at $k = 2, 10, 32$ on MNIST); three-layer ones reconstruct better while ceasing to be projections (defect 0.28–0.76, cycle defect 0.10–0.20). They send each training point back to itself without fixing their decoded manifold pointwise. Across all configurations the idempotence defect is close to an increasing function of ρ (Figure 2, right).

Table 4: Is $Z_\perp$ needed for a gain over PCA? Held-out reconstruction gain over PCA at the same k (4000 training / 1000 test images, same decoder architecture k–256–784, tanh, JAX implementation, 100 epochs). $\mathcal{D}(Z)$: end-to-end autoencoder on its own code. $\mathcal{D}(Z_\parallel)$: the same trained decoder fed only the realizable part of its code (an affine map of the pixels fitted on the training split). $\mathcal{D}(Z_\perp)$: the same decoder fed only the remainder, placed at the training mean of the code. "Decoder on PCA scores": a decoder trained on the PCA-k scores, a code with $Z_\perp = 0$ by construction.

				held-out gain over PCA			
Data	k	PCA test	ρ (AE)	$\mathcal{D}(Z)$	$\mathcal{D}(Z_\parallel)$	$\mathcal{D}(Z_\perp)$	decoder on PCA scores
MNIST	2	0.912	0.311	+0.065	−0.073	−0.230	+0.025
	10	0.713	0.155	+0.097	+0.044	−0.333	+0.095
	32	0.502	0.042	+0.013	−0.029	−0.524	+0.047
Fashion-MNIST	2	0.734	0.159	+0.094	+0.050	−0.516	+0.064
	10	0.529	0.094	+0.058	+0.052	−0.496	+0.058
	32	0.419	0.040	+0.010	+0.012	−0.586	+0.023

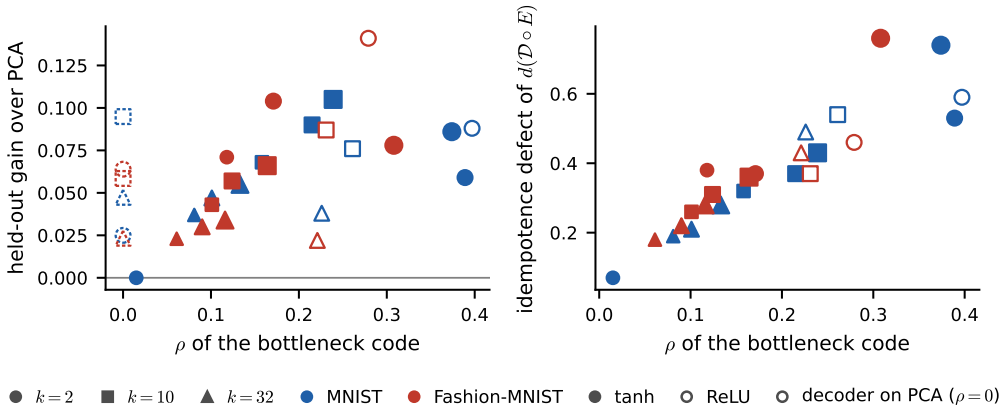

Figure 2: Nonlinear autoencoder sweep. **Left:** held-out reconstruction gain over PCA against ρ of the bottleneck code. In end-to-end training every configuration with $\rho > 0.05$ gains, and for tanh at $k \geq 10$ the gain grows with ρ; at $k = 2$ and for ReLU the two decouple. Dotted hollow markers at $\rho = 0$: a nonlinear decoder trained on the PCA scores gains too, so $Z_\perp$ is not necessary (Table 4). **Right:** idempotence defect of $d(\mathcal{D} \circ E)$ against ρ: shallow, near-linear autoencoders are projections (Proposition 6); deeper ones reconstruct better while ceasing to be projections, i.e. they recover every training point without recovering the manifold as a fixed set.

7 Local and Layer-wise Realizability

Local realizability (Appendix I) replaces the global projection by one per K-neighborhood, on its d leading local principal coordinates. For a smooth, non-degenerate function of a C^2 manifold the local residual vanishes as the neighborhood shrinks (Proposition 7), and it is identically zero for $K \leq d+1$. On a Swiss roll the true coordinates, t-SNE and UMAP all become locally realizable at small K (curved but smooth), while on MNIST and Fashion-MNIST t-SNE and UMAP stay 0.07–0.36 above PCA's reference level at every scale tested: neighborhood embeddings of images are not smooth functions of pixel neighborhoods, and their remainder is not curvature.

Layer-wise realizability (Appendix J) applies the decomposition between consecutive layers of a trained width-128 ReLU MLP. Hidden layers of equal width are not projections of one another ($\rho(H_{\ell-1}, H_\ell)$ from 0.14–0.19 at the first $128 \to 128$ transition down to 0.09–0.10 at the fourth), while the representation moves steadily away from any projection of the input ($\rho(X, H_\ell)$ up to 0.28–0.33) and the participation ratio follows its own schedule; geometry changes at every layer with dimension changing at none. Replacing a layer by its realizable part changes the logits by at most $L_{>\ell} \left\| h_\ell - \tilde{h}_\ell \right\|$, which gives a per-sample certificate against the margin (Proposition 8). Empirically,

linearizing the first layer costs 2.4–3.6 accuracy points and any later layer at most 0.8; because ρ, $L_{>\ell}$ and the distance to the logits all fall with depth in this network, we claim the composite certificate, not that ρ alone ranks layers.

8 Related Work

Chun et al. (2026) measure how many dimensions a representation has; we measure how much of it is a projection, and Proposition 3 is the same finite-sample failure in our quantity. Moisescu-Pareja et al. (2026) study the manifolds representations form; we study the map that produced them. I-Con (Alshammari et al., 2025) unifies objectives as KL minimization, and its invariance group fixes our coordinate class; Uscidda et al. (2025) regularize map distortion, of which (2) is the exact linear case. Krzanowski (2005) distinguishes axes that are linear combinations of the variables from subspaces found directly, the distinction Theorem 2 makes exact; linear autoencoders recover the principal subspace but not its axes (Baldi & Hornik, 1989; Plaut, 2018). Spectral and neighborhood methods (Belkin & Niyogi, 2003; Coifman & Lafon, 2006; Lee & Verleysen, 2011; Amid & Warmuth, 2019; Wang et al., 2021) are cross-space embeddings in our sense. Regression, CCA and shape indices (Raghu et al., 2017; Kornblith et al., 2019; Williams et al., 2021; Schönemann, 1966) compare representations up to a transformation class; the linear-regression index is already asymmetric and equals $1 - \rho^2$, so what we add is the realization theorem and the separate evaluation of $Z_{\|}$ and $Z_{\perp}$. Appendix K gives a longer discussion.

9 Limitations and Conclusion

Global realizability is the wrong criterion for curved data, which the local version addresses at the price of a neighborhood size and a local dimension. ρ is uninformative for $n \lesssim D$, and for feature sources with $p \gtrsim n$, which is why rich sources are evaluated out of sample. The reflection check of Proposition 6(iii) is conclusive only at points fixed by the symmetry. Our autoencoder conclusions concern MLP autoencoders trained with squared loss (plain, contractive, variational); our networks are small; our embeddings are t-SNE, UMAP and PaCMAP; and our sources are raw pixels. Appendix L discusses each.

Fewer coordinates is not an operation. Every linear reduction is a projection inside the original space, read through coordinates that may distort geometry but never leave it. For any finite centered representation Z of a declared source X, including one produced by a nonlinear map, the representation splits exactly into a projection-realizable part and an orthogonal remainder. For neighborhood embeddings the remainder is small in energy and carries their neighborhood fidelity beyond PCA; for autoencoders it describes the code that end-to-end training chose rather than a requirement of reconstruction. The question to ask of a representation is therefore not only *how many dimensions remain* but *how much of it is a projection of its source, and what the rest is doing*.

Ethics statement

We have read the ICLR Code of Ethics and find that this work does not raise significant ethical concerns. It is a methodological study of representation geometry that uses only public benchmark image datasets (MNIST, Fashion-MNIST and CIFAR-10). It involves no human subjects, no personal or sensitive data, and no application with a direct potential for harm. In keeping with transparency and research integrity, we release all code and report every experiment we ran for the claims made, including the negative result that the remainder $Z_{\perp}$ is not required for an autoencoder's gain over PCA.

Reproducibility statement

We have made every effort to make our results reproducible. All proofs are in Appendix A, and Appendix N details the datasets, model configurations, hyperparameters and training procedures. The complete code is provided in the supplementary material: eight experiment scripts over one shared module `common.py` and a small JAX module `jax_ae.py`, with `make_tables.py` emitting every table and figure and `run_all.sh` running the whole pipeline in dependency order. All random seeds are fixed, `common.shared_indices` replays the subsampling so that every table

uses the same images, and the README lists the software dependencies and maps each script to the tables it produces. All data are public (the MNIST, Fashion-MNIST and CIFAR-10 test sets), and the full pipeline runs in about three hours on two CPU cores. The code will be released publicly upon acceptance.

USE OF LARGE LANGUAGE MODELS

We used a large language model (Claude, Anthropic) as an assistant. Its main use was language revision: improving wording and clarity and condensing the manuscript to the page limit.

REFERENCES

Matteo Alleman, Jack Lindsey, and Stefano Fusi. Task structure and nonlinearity jointly determine learned representational geometry. In *International Conference on Learning Representations (ICLR)*, 2024.

Shaden Alshammari, John Hershey, Axel Feldmann, William T. Freeman, and Mark Hamilton. I-Con: A unifying framework for representation learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Ehsan Amid and Manfred K. Warmuth. TriMap: Large-scale dimensionality reduction using triplets. *arXiv preprint arXiv:1910.00204*, 2019.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59(4–5):291–294, 1988.

Chanwoo Chun, Abdulkadir Canatar, SueYeon Chung, and Daniel D. Lee. Estimating dimensionality of neural representations from finite samples. In *International Conference on Learning Representations (ICLR)*, 2026.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

Danish Ebadulla, Aditya Gulati, and Ambuj Singh. Normalized space alignment: A versatile metric for representation analysis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025.

Mateus Espadoto, Rafael M. Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, 2021.

Marco Fumero, Marco Pegoraro, Valentino Maiorca, Francesco Locatello, and Emanuele Rodolà. Latent functional maps: a spectral framework for representation alignment. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Mehmet Gönen and Ethem Alpaydın. Supervised learning of local projection kernels. *Neurocomputing*, 73(10–12):1694–1703, 2010.

Sharut Gupta, Joshua Robinson, Derek Lim, Soledad Villar, and Stefanie Jegelka. Structuring representation geometry with rotationally equivariant contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Geoffrey E. Hinton and Sam T. Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15, 2002.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Alan Julian Izenman. Introduction to manifold learning. *WIREs Computational Statistics*, 4(5): 439–446, 2012.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065):20150202, 2016.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Wojtek J. Krzanowski. Axes in multivariate analysis. In *Encyclopedia of Statistics in Behavioral Science*. Wiley, 2005. doi:10.1002/0470013192.bsa035.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

John A. Lee and Michel Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. *Procedia Computer Science*, 4:538–547, 2011.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Marina Meilă and Hanyu Zhang. Manifold learning: What, how, and why. *Annual Review of Statistics and Its Application*, 11:393–417, 2024.

Gabriela Moisescu-Pareja, Gavin McCracken, Harley Wiltzer, Colin Daniels, Vincent Létourneau, Doina Precup, and Jonathan Love. On the geometry and topology of representations: The manifolds of modular addition. In *International Conference on Learning Representations (ICLR)*, 2026.

Elad Plaut. From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253*, 2018.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 833–840, 2011.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Dan Shiebler. Functorial manifold learning. In *Proceedings of the Fourth International Conference on Applied Category Theory (ACT 2021)*, volume 372 of *EPTCS*, pp. 1–13, 2022.

Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J. Maybank. Geometric mean for subspace selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):260–274, 2009.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Théo Uscidda, Luca Eyring, Karsten Roth, Fabian J. Theis, Zeynep Akata, and Marco Cuturi. Disentangled representation learning with the Gromov-Monge gap. In *International Conference on Learning Representations (ICLR)*, 2025.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks – ICANN 2001*, pp. 485–491, 2001.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Ke Ye and Lek-Heng Lim. Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197, 2016.

A Proofs and Remarks

Theorem 1. Write $s_1 = U_k a$, $s_2 = U_k b$; then $c(s_1) = a$, $c(s_2) = b$, and $\|s_1 - s_2\|^2 = (a - b)^\top U_k^\top U_k (a - b) = \|a - b\|^2$; inner products likewise.

Theorem 2. Let the columns of U be an orthonormal basis of $\mathrm{row}(W)$. Each row of W lies in $\mathrm{span}(U)$, so $W = AU^\top$ for some $A \in \mathbb{R}^{k\times k}$; $\mathrm{rank}\, W = \mathrm{rank}\, U^\top = k$ forces $\mathrm{rank}\, A = k$. Then $z = AU^\top x = AU^\top(UU^\top x) = AU^\top \hat{x}$ since $U^\top U = I_k$, and $\hat{x} = UU^\top x = UA^{-1}z$. The thin QR $W^\top = QR$ gives $W = R^\top Q^\top$ with Q orthonormal and $R^\top$ invertible.

Corollary 1. If $W = U^\top$ with $U^\top U = I_k$ then $WW^\top = I_k$; conversely $WW^\top = I_k$ makes the rows of W orthonormal, so $U = W^\top$ and $P_S = UU^\top = W^\top W$.

Equations (1)**–**(2)**.** $z_i - z_j = A(c_i - c_j)$ and $\|c_i - c_j\| = \|\hat{x}_i - \hat{x}_j\|$, so the ratio is $\|Av\| \,/\, \|v\| \in [\sigma_{\min}(A), \sigma_{\max}(A)]$, with both endpoints attained; and $(z_i - z_j)^\top A^{-\top} A^{-1} (z_i - z_j) = \|c_i - c_j\|^2$.

Theorem 3. $\mathcal{D}'(E'(x)) = \mathcal{D}(h^{-1}(h(E(x)))) = \mathcal{D}(E(x))$; if h is not an isometry, $\|h(z_i) - h(z_j)\| \neq \|z_i - z_j\|$ in general.

$\rho_{\mathrm{lin}} \leq \rho_{\mathrm{sim}}$**.** $\{sQ : Q^\top Q = I_k, s > 0\}$ consists of rank-k matrices, hence is contained in the feasible set of Proposition 1(i).

Expected null. The null is a random representation, not a permutation test: no rows or labels are shuffled. Let Z have i.i.d. $\mathcal{N}(0, \sigma^2)$ entries independent of X, let $C = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $r = \operatorname{rank} X$. Since X is centered, $\operatorname{col}(X)$ is an r-dimensional subspace of $\mathbf{1}^\perp = \operatorname{col}(C)$, and each column of CZ is an isotropic Gaussian in that $(n-1)$-dimensional space. Hence $\|Z_\parallel\|_F^2 / \sigma^2 \sim \chi^2_{kr}$ and $\|Z_\perp\|_F^2 / \sigma^2 \sim \chi^2_{k(n-1-r)}$ are independent, $\rho^2_{\text{lin}} \sim \text{Beta}\big(\frac{k(n-1-r)}{2}, \frac{kr}{2}\big)$, and $\mathbb{E}\,\rho^2_{\text{lin}} = 1 - r/(n-1)$ exactly, with variance $O(1/(nk))$. For non-Gaussian i.i.d. entries the same expression holds for the ratio of expectations, $\mathbb{E}\,\|Z_\perp\|_F^2 / \mathbb{E}\,\|Z\|_F^2$. With $n = 5000$ and $\operatorname{rank} X = 642$ (MNIST) or 784 (Fashion-MNIST) this gives $\rho^2_{\text{lin}} \approx 0.872$ and 0.843, i.e. $\rho_{\text{lin}} \approx 0.934$ and 0.918, matching the observed residuals 0.934 and 0.918.

Genericity in Proposition 3. For $n \le D + 1$ the centered matrix $X \in \mathbb{R}^{n \times D}$ has $D \ge n - 1$ columns, each lying in the $(n-1)$-dimensional space $\mathbf{1}^\perp \subset \mathbb{R}^n$; its rank is $n - 1$ unless those columns fail to span $\mathbf{1}^\perp$, a null event for absolutely continuous data.

Strictness in Proposition 2. Take $X = (1, -1, 0)^\top$ and $Z = X + e$ with $e = (1, 1, -2)^\top \perp X$, so $Z_\parallel = X$, $Z_\perp = e$. The best linear reconstruction of X from Z has squared error 1.5; from $Z_\parallel$ it is 0. In 200 random trials with a $\tanh$ encoder plus noise ($n = 200$, $D = 20$, $k = 3$) the chain (5) held in every case.

Kernel maps. A feature map $\phi : \mathbb{R}^D \to \mathcal{H}$ is a cross-space embedding that typically *increases* dimension, and kernel PCA then projects inside $\mathcal{H}$ (Schölkopf et al., 1998). Kernel PCA scores are exactly realizable relative to $\mathcal{H}$ and in general not relative to $\mathbb{R}^D$; realizability is always relative to a declared source, which is what relative realizability in §4 formalizes.

Proposition 1. $\{UA^\top : U^\top U = I_k,\ A \text{ invertible}\}$ is the set of rank-k $D \times k$ matrices (thin QR gives one inclusion, the converse is immediate), which is dense in $\mathbb{R}^{D \times k}$ but not closed; so the infimum of the continuous objective over it equals the unconstrained least-squares minimum $\min_M \|Z - XM\|_F$, attained at $M = X^+Z$ with fitted values $\Pi_X Z$. It is attained inside the constrained class iff some minimizer has rank k. The minimizers are $X^+Z + N$ with $XN = 0$; since $\operatorname{col}(X^+Z) \subseteq \operatorname{row}(X) \perp \ker X$, adding N with columns in $\ker X$ raises the rank by up to $\dim \ker X$, so a rank-k minimizer exists iff $\operatorname{rank}(X^+Z) + \dim \ker X \ge k$; for full column rank $\ker X = 0$ and the condition is $\operatorname{rank}(X^+Z) = k$. The same orthogonality gives $\|X^+Z + N\|_F^2 = \|X^+Z\|_F^2 + \|N\|_F^2$, so X^+Z is the minimum-norm minimizer. Orthogonality of the residual to $\operatorname{col}(X)$ is the normal equation. $R^2 = \|Z_\parallel\|_F^2 / \|Z\|_F^2$ by Pythagoras, giving (4). (iv) follows from (i).

Proposition 2. $e(Z)$ is the squared distance from X to $\operatorname{col}(Z) \subseteq \operatorname{col}(Z_\parallel) \oplus \operatorname{col}(Z_\perp)$. Since $Z_\perp \perp \operatorname{col}(X)$ and every column of X lies in $\operatorname{col}(X)$, the orthogonal projection of X onto $\operatorname{col}(Z_\parallel) \oplus \operatorname{col}(Z_\perp)$ equals its projection onto $\operatorname{col}(Z_\parallel)$, so the distance from X to the larger space equals $e(Z_\parallel)$, and $e(Z) \ge e(Z_\parallel)$ because $\operatorname{col}(Z)$ is a subspace of it. For the second inequality, $Z_\parallel W = XMW$ is a linear map of X of rank at most k, whose reconstruction error is at least the PCA-k error by Eckart–Young. Both inequalities are strict in general (a three-point example in Appendix A), and the last statement is the classical result of Bourlard & Kamp (1988) in the present notation.

Proposition 3. Centered Z and $\operatorname{col}(X)$ both lie in $\mathbf{1}^\perp \subset \mathbb{R}^n$, of dimension $n-1$; if $\operatorname{rank} X = n-1$, $\operatorname{col}(X) = \mathbf{1}^\perp$ and $\Pi_X Z = Z$.

Proposition 4 (full statement). Let $\mathcal{L}(\theta) = \sum_i d\big(x_i, \mathcal{D}_\theta(E_\theta(x_i))\big)$ with d a pointwise discrepancy on $\mathbb{R}^D$ that vanishes only on the diagonal. Then (i) any latent reparameterization h leaves $\mathcal{L}$ and the reconstruction map $\mathcal{D} \circ E$ unchanged; (ii) for any non-identity isometry g of $\mathbb{R}^D$ that does not fix the data set pointwise, the reconstruction $g \circ \mathcal{D} \circ E$ has strictly larger loss than $\mathcal{D} \circ E$ whenever the latter is exact on the data. If instead $\mathcal{L}$ depends on the outputs only through a statistic $T(\{\mathcal{D}(E(x_i))\})$ that is invariant under a group $\mathcal{G}$ acting on $\mathbb{R}^D$, then $g \circ \mathcal{D} \circ E$ has the same loss for every $g \in \mathcal{G}$.

Proof. (i) is Theorem 3. (ii) If $\mathcal{D}(E(x_i)) = x_i$ for all i then $d(x_i, g(x_i)) > 0$ for some i by hypothesis, so the loss of $g \circ \mathcal{D} \circ E$ is positive while that of $\mathcal{D} \circ E$ is zero. The last statement is immediate from the invariance of T.

Proposition 5 (full statement). Let P be the data distribution on $\mathbb{R}^D$ and d a pointwise discrepancy with $d(x, y) = 0$ iff $x = y$. A reconstruction map $R = \mathcal{D} \circ E$ minimizes $\mathbb{E}_P\, d(x, R(x))$ if and only if $R(x) = x$ for P-almost every x. Consequently, for every measurable bijection φ of $\mathrm{supp}(P)$ with $\varphi \neq \mathrm{id}$ on a set of positive P-measure, in particular every non-trivial isometry, every involution, and every map preserving all pairwise distances and neighborhood relations, $R = \varphi$ has strictly positive loss. This holds even when P is itself invariant under φ, because the loss pairs each sample with its own reconstruction and not with the distribution. On a finite sample the same statement holds with "P-a.e." replaced by "on the training set".

Proof. $\mathbb{E}_P\, d(x, R(x)) \geq 0$ with equality iff $d(x, R(x)) = 0$ P-a.s., iff $R(x) = x$ P-a.s. If $\varphi \neq \mathrm{id}$ on a set B with $P(B) > 0$ then $d(x, \varphi(x)) > 0$ on B and the expectation is positive.

Proposition 6 (full statement). Let $R = \mathcal{D} \circ E$ be C^1 and let $M_{\mathcal{D}} = \mathcal{D}(\mathcal{Z})$ be the decoded manifold, so that $R(\mathbb{R}^D) \subseteq M_{\mathcal{D}}$ by construction. (i) If $E \circ \mathcal{D} = \mathrm{id}$ on $E(\mathbb{R}^D)$ (latent cycle-consistency) then $R \circ R = R$; conversely, if $\mathcal{D}$ is injective on $E(\mathbb{R}^D)$, then $R \circ R = R$ implies $E \circ \mathcal{D} = \mathrm{id}$ on $E(\mathbb{R}^D)$. (ii) If $R \circ R = R$, then at every reconstructed point $x^\star = R(x)$ the Jacobian $J = dR_{x^\star}$ satisfies $J^2 = J$, its spectrum is contained in $\{0, 1\}$, and $\mathrm{rank}\, J = \mathrm{tr}\, J$. If moreover $\mathcal{D}$ is an immersion at $z^\star = E(x^\star)$ and $dE_{x^\star}$ has rank k, then $\mathrm{rank}\, J = \mathrm{tr}\, J = k$ and $\mathrm{col}(J) = T_{x^\star} M_{\mathcal{D}}$: R is locally an (oblique) projection onto the decoded manifold. (iii) Suppose instead $R = g \circ R_0$ with R_0 a projection as in (ii) and g a non-trivial isometry of $\mathbb{R}^D$ with $g(M_{\mathcal{D}}) = M_{\mathcal{D}}$. At a point $x^\star = R_0(x^\star)$ that g fixes, dg maps $T_{x^\star} M_{\mathcal{D}}$ to itself, and $dR_{x^\star}$ has an eigenvalue of modulus one different from 1 (-1 for a reflection, complex for a rotation) unless dg is the identity on that tangent space. Where $g(x^\star) \neq x^\star$ on a curved $M_{\mathcal{D}}$, dg carries $T_{x^\star} M_{\mathcal{D}}$ to the different subspace $T_{g(x^\star)} M_{\mathcal{D}}$ and no such eigenvalue need exist.

Proof. (i) $R(R(x)) = \mathcal{D}\big(E(\mathcal{D}(z))\big)$ with $z = E(x)$. If $E(\mathcal{D}(z)) = z$ this equals $\mathcal{D}(z) = R(x)$. Conversely, $R \circ R = R$ says $\mathcal{D}(E(\mathcal{D}(z))) = \mathcal{D}(z)$ for all $z \in E(\mathbb{R}^D)$; both $E(\mathcal{D}(z))$ and z lie in $E(\mathbb{R}^D)$, so injectivity of $\mathcal{D}$ there gives $E(\mathcal{D}(z)) = z$. (ii) Differentiating $R \circ R = R$ at $x^\star = R(x^\star)$ gives $J^2 = J$; an idempotent has spectrum in $\{0, 1\}$ and rank equal to trace. $J = d\mathcal{D}_{z^\star}\, dE_{x^\star}$ and $\mathrm{col}(J) \subseteq T_{x^\star} M_{\mathcal{D}}$ since R maps into $M_{\mathcal{D}}$; when $d\mathcal{D}_{z^\star}$ is injective ($\mathcal{D}$ an immersion) and $dE_{x^\star}$ is onto $\mathbb{R}^k$, $\mathrm{rank}\, J = k = \dim T_{x^\star} M_{\mathcal{D}}$, so the inclusion is an equality. Without these conditions $\mathrm{rank}\, J < k$ is possible (a constant map is idempotent). (iii) At a fixed point, $g(M_{\mathcal{D}}) = M_{\mathcal{D}}$ and $g(x^\star) = x^\star$ give $dg\, T_{x^\star} M_{\mathcal{D}} = T_{x^\star} M_{\mathcal{D}}$, so $T := T_{x^\star} M_{\mathcal{D}} = \mathrm{col}(J_0)$ is invariant under $dR = dg\, J_0$, on which dR acts as the orthogonal map $dg|_T$; an orthogonal map other than the identity has an eigenvalue of modulus one other than 1. Away from fixed points the invariance fails: for the unit circle with radial projection R_0 and the reflection $g(x_1, x_2) = (x_1, -x_2)$, at $x^\star = (1, 1)/\sqrt{2}$ the matrix $dg\, J_0$ is nilpotent.

B A Worked Example of the Three Operations

A worked example fixes the three objects. Let $x = (3, 2)^\top$. A basis change by the rotation U with columns $(1, 1)^\top/\sqrt{2}$ and $(-1, 1)^\top/\sqrt{2}$ reports $U^\top x = (5, -1)^\top/\sqrt{2}$: two numbers for the same vector. Let S be the horizontal axis with $u = (1, 0)^\top$. Then $\hat{x} = (3, 0)^\top$ stays in $\mathbb{R}^2$, and $z = u^\top x = 3$ is its coordinate. If an embedding algorithm instead assigns the sample $y = 1.7$ because that location best serves a neighborhood objective, y is a target-space location that need not be the coordinate of any projection of x. The question this paper answers is how to tell, for a learned f, which case one is in, and how much of f is each.

C Comparison with Representation-Similarity Indices

We say this plainly, because a reader will otherwise say it for us: ρ_{lin} is the residual of a multivariate linear regression, and $1 - \rho^2_{\mathrm{lin}}$ is its R^2. It is the same quantity as the linear-regression similarity index of Kornblith et al. (2019) and the accuracy of a linear probe with Z as target, and it is closely related to CCA-based measures (Raghu et al., 2017). What the decomposition adds is not the number

Table 5: ρ against representation-similarity indices for the $k = 2$ representations of §6.1 (source: pixels). ρ_{sim}: similarity residual (Procrustes with scaling onto a k-dimensional projection of X, Appendix M). CKA: linear centered kernel alignment between X and Z. CCA_{50} / $\mathrm{CCA}_{\mathrm{full}}$: mean canonical correlation after truncating X to its 50 leading, or all 642/784 nonzero, singular directions. Procrustes: angular orthogonal-Procrustes shape distance between X and zero-padded Z (Williams et al., 2021) ($0°$ = identical shape, $90°$ = unrelated).

Data	Method	ρ	ρ_{sim}	CKA	CCA_{50}	$\mathrm{CCA}_{\mathrm{full}}$	Procrustes
MNIST	PCA	$< 10^{-14}$	$< 10^{-14}$	0.68	1.00	1.00	65°
	Random orthonormal proj.	$< 10^{-14}$	0.003	0.20	0.88	1.00	77°
	Linear autoencoder	$< 10^{-14}$	0.000	0.68	1.00	1.00	65°
	Nonlinear autoencoder	0.011	0.011	0.68	1.00	1.00	65°
	t-SNE	0.409	0.410	0.45	0.88	0.91	70°
	UMAP	0.420	0.422	0.39	0.86	0.90	71°
	PaCMAP	0.427	0.427	0.42	0.87	0.90	71°
	null (random Z)	0.934	0.937	0.00	0.10	0.36	89°
Fashion-MNIST	PCA	$< 10^{-14}$	0.000	0.96	1.00	1.00	47°
	Random orthonormal proj.	$< 10^{-14}$	$< 10^{-14}$	0.23	0.87	1.00	70°
	Linear autoencoder	$< 10^{-14}$	0.001	0.95	1.00	1.00	47°
	Nonlinear autoencoder	0.157	0.158	0.81	0.98	0.99	51°
	t-SNE	0.227	0.228	0.74	0.96	0.97	53°
	UMAP	0.227	0.231	0.65	0.96	0.97	56°
	PaCMAP	0.243	0.244	0.68	0.95	0.97	55°
	null (random Z)	0.918	0.918	0.00	0.10	0.40	89°

but the objects and their reading: Theorem 2 says that the fitted part $Z_{\parallel}$ is, exactly, the coordinates of one fixed orthogonal projection of the source, so ρ_{lin} measures the fraction of a representation that is not a projection; the two parts $Z_{\parallel}$ and $Z_{\perp}$ can each be evaluated as representations, which a similarity score cannot; the finite-sample and null statements below say when the number means anything; and, in §5 and §I, the same construction yields certificates and a local version that have no counterpart among similarity indices. Table 5 makes the difference concrete on the representations of §6.1: linear CKA (Kornblith et al., 2019) assigns 0.68 to PCA and 0.20 to a random orthonormal projection of the same data although both are exactly realizable ($\rho < 10^{-14}$), because CKA measures how much of the source *variance* the representation captures, not whether it is a projection; the mean canonical correlation at full rank assigns 1.00 to every linear map, as it should, but also 0.36 to pure noise at $n = 5000$, $\operatorname{rank} X = 642$, the finite-sample inflation that Proposition 3 makes explicit and that our null accounts for. The orthogonal-Procrustes shape distance (Williams et al., 2021) behaves like CKA ($65°$ for PCA, $77°$ for the random projection and $70°$ for t-SNE on MNIST): as a symmetric distance invariant to rotation and scale it compares how the two sides distribute variance, so an exact projection that keeps little variance looks as far from the source as a t-SNE layout. The similarity residual ρ_{sim}, which is Procrustes with isotropic scaling (Schönemann, 1966) restricted to a k-dimensional projection of the source, stays within 0.004 of ρ_{lin} for every representation in the table: the neighborhood layouts are realizable up to rotation and scale exactly as far as they are realizable up to an arbitrary linear coordinate change. None of these indices returns the parts $Z_{\parallel}$ and $Z_{\perp}$ whose separate evaluation is the content of §6.1.

D Linear Autoencoders and the Recovery Checks in Detail

For a *linear* autoencoder the encoder $W_e \in \mathbb{R}^{k \times D}$ factors by Theorem 2 as $AU^{\top}$, and what is geometrically determined is exactly $S = \operatorname{span}(U)$ and the metric G of (2); the latent axes are determined only up to A. With squared loss, S at the global optimum is the principal subspace (Baldi & Hornik, 1989; Plaut, 2018), so the latent code is $z = A \cdot$ (PCA scores): the same projection as PCA read through an invertible coordinate change, which is a rotation of the principal components only when $A^{\top}A \propto I$. In our experiments (§6) the trained linear autoencoders have $\rho_{\mathrm{sim}} \approx 0$ and $\kappa(A) \leq 1.6$, so the optimizer happened to land near a similarity of the PCA scores; that is a property of the solution found, not of the architecture. Which weights reveal the principal subspace depends on whether the trained map $W_d W_e$ is an orthogonal or an oblique projector. The subspace projected *onto* is always $\operatorname{col}(W_d)$, so it is recovered by an orthonormal basis of the decoder's column space

(thin QR or SVD of W_d); Table 6 verifies it to within $1°$–$3°$, except on MNIST at $k = 10$, where one direction swaps with the nearly degenerate eleventh principal component ($85°$, §6.1). The encoder's row space $\mathrm{row}(W_e)$ equals it only when the projector is orthogonal; for the oblique projectors found here it sits $41°$–$49°$ away (§6.1), and the encoder weights are oblique measurement coordinates, not a basis of S.

For a *nonlinear* autoencoder there is no basis to know, in two senses. The latent space is not a subspace of the input space, so no input directions "are" the latent axes; locally the row space of the encoder Jacobian $J_E(x)$ and the column space of the decoder Jacobian $J_{\mathcal{D}}(z)$ play the role of U, but these frames move with x, which is what having no global source subspace means. And within the latent space, Theorem 3 makes axes, distances and angles unidentified unless a symmetry-breaking term is added (an isotropic prior, a Jacobian or orthogonality penalty, a post-hoc PCA of the codes, which rotates *inside* the latent space and has no fixed correspondence to input directions). What can be said is quantitative: $Z_{\|}$ is the part of the code that is a fixed projection of the input, with recoverable basis $U = \mathrm{orth}(X^+ Z)$, and $Z_{\perp}$ is the part that has no basis in the input space by construction.

The reassuring reading: the output space is fixed by construction (the decoder emits D numbers), and its coordinate labels are fixed by the loss, which compares those numbers entry by entry with the original x. The symmetry one senses is real but lives entirely in the latent and cancels before the output; in the linear case, at the global optimum of the squared loss, $W_d W_e = UU^\top$ regardless of which A the optimizer chose.

Two consequences deserve to be spelled out. First, the decoder is a *left inverse of the encoder on the image of the data only*: $\mathcal{D} = E^{-1}$ on $E(\mathrm{supp}\, P)$, and nowhere else is $\mathcal{D}$ constrained; "the inverse function" is the right description of $\mathcal{D}$ exactly on that set. Second, "returning to another space with the same values" is not even in the hypothesis class: the loss is evaluated in the coordinates of x, so a bijective copy $Y = \varphi(\mathcal{X})$ in any other space would have to be compared through φ^{-1}, which the loss does not apply. Between training samples, however, R is fixed by inductive bias rather than by the loss, so the operational certificate of Proposition 5 is the *held-out* reconstruction residual, which we report in Table 10.

The linear case is the sanity check: $R(x) = W_d W_e x$, $J = W_d W_e$ constant, and Proposition 6(ii) is the idempotence test of Table 6, with symmetry of J the additional statement that the projection is orthogonal. For a nonlinear autoencoder we evaluate dR analytically at data points and report the idempotence defect $\left\|J^2 - J\right\| / \|J\|$, the number of eigenvalues near 1 (predicted k), the number of eigenvalues of modulus $> \frac{1}{2}$ with negative real part (predicted 0; a reflection would give ≥ 1 at the points it fixes, but part (iii) says nothing about the points it moves, so this count is a necessary check rather than a certificate), and the map-level defect $\|R(R(x)) - R(x)\| / \|R(x)\|$ on held-out data (Table 10). Note that reconstruction loss alone does not enforce (i): a decoder can reconstruct the training set exactly while $E \circ \mathcal{D} \neq \mathrm{id}$ off the latent image, in which case R is not a projection and its Jacobian spectrum is not in $\{0, 1\}$; the experiments below show this happens at larger depth, and it is the sense in which such an autoencoder has *not* recovered the original space as a fixed set even while recovering every original point.

The remaining checks are geometric rather than nominal: for a linear autoencoder, that $W_d W_e$ is symmetric and idempotent and that $\mathrm{col}(W_d)$ is the principal subspace; for a nonlinear one, that the residual $r(x) = x - \mathcal{D}(E(x))$ is orthogonal to the decoded manifold's tangent plane $\mathrm{col}(J_{\mathcal{D}}(z))$ (first-order optimality) and that this tangent plane lies inside the data's local tangent space. Table 6 (Appendix D) reports both.

The caveat that makes the question non-trivial is the last clause of Proposition 4. The moment the objective is distribution-level (adversarial, Wasserstein or Gromov–Monge, or a neighborhood objective such as SNE's KL), the loss is invariant to every transformation preserving the compared statistic, and a reflected or rotated copy of the data is a perfectly good output. That is exactly the freedom Uscidda et al. (2025) regularize, and it is why the coordinate class in Definition 1 must be declared per objective: pointwise reconstruction pins the output space and leaves only the latent free; distribution matching leaves the output free up to the symmetry group of the loss, so for SNE-type objectives ρ_{sim}, not ρ_{lin}, is the residual that respects the objective's own invariance (Alshammari et al., 2025).

Findings on MNIST and Fashion-MNIST. **(i) Linear methods are exactly realizable** ($\rho_{\mathrm{lin}} < 10^{-14}$) with $Z = Z_{\|}$, as Theorem 2 requires; the random projection shares PCA's residual and none of its quality, so ρ measures *whether* a representation is a projection, not whether it is a good one. **(ii) The linear autoencoder is PCA in oblique coordinates.** Its residual is zero, its realized subspace is within 1°–2° of the PCA subspace in score space at $k = 2$ (mean angle; at $k = 10$ on MNIST the mean is 10° with one 85° direction, a swap between the nearly degenerate tenth and eleventh principal components that costs nothing in reconstruction), its reconstruction error equals PCA's to three decimals, and Table 6 shows $W_d W_e$ idempotent to 3×10^{-4} but visibly non-symmetric (0.02–0.06): the trained network is an *oblique* projector onto the principal subspace, which is why its encoder row space, measured without data weighting, sits 41°–49° from the PCA subspace while its codes are indistinguishable from PCA scores. **(iii) Neighborhood embeddings are mostly projection by energy, not by neighborhood fidelity.** t-SNE and UMAP have $\rho = 0.41$–0.42 on MNIST and 0.23–0.24 on Fashion-MNIST, far below the null (0.93, 0.92) and far above zero. By energy, $1 - \rho^2$, their realizable parts carry 82–83% (MNIST) and 94–95% (Fashion-MNIST) of the layout. Two utilities of those parts must be kept apart. Their *neighborhood fidelity* is PCA's: trustworthiness 0.76–0.77 against PCA's 0.75 on MNIST and 0.90–0.92 against 0.91 on Fashion-MNIST, while the full layouts reach 0.96–0.98 and 0.98–0.99. Their *class separability* is above PCA's, 10-NN accuracy 0.58–0.62 against 0.44 on MNIST and 0.63–0.64 against 0.52 on Fashion-MNIST, because the projection they realize is not PCA's (mean principal angle 26°–36° in score space) and is more discriminative; but it remains far below the full layouts' 0.93–0.94 and 0.75–0.78. At $k = 2$ the remainder $Z_\perp$ on its own is near chance (trustworthiness ≈ 0.5, kNN 0.15–0.23); at $k = 10$ it is not (UMAP's $Z_\perp$ alone reaches kNN 0.43 on MNIST and 0.29 on Fashion-MNIST, against nulls of 0.10 and 0.11). At $k = 2$ the neighborhood fidelity is therefore in neither part alone: a small non-realizable component, useless by itself, is what turns a PCA-grade projection into a t-SNE-grade layout, and it is lost the moment $Z_\perp$ is removed. **(iv) Autoencoders: the gain over PCA comes with $Z_\perp$, but does not need it.** At $k = 2$ on MNIST the $\tanh$ autoencoder has $\rho = 0.011$ and PCA's reconstruction error (0.907 vs. 0.907): at this width and budget it found a near-linear solution, which ρ detects without inspecting weights. At $k = 10$ on MNIST ($\rho = 0.21$) and on Fashion-MNIST ($\rho = 0.16$ at $k = 2$, 0.12 at $k = 10$) it improves reconstruction over PCA (0.590 vs. 0.708 at $k = 10$ on MNIST; on Fashion-MNIST 0.641 vs. 0.728 at $k = 2$ and 0.460 vs. 0.528 at $k = 10$), and its realized subspace drifts 16°–19° from PCA's, whereas at $\rho = 0.011$ it sat within 1.5°. This co-occurrence is not a requirement: §6.2 trains the same nonlinear decoder on the PCA scores, a code with $Z_\perp = 0$ by construction, and obtains a comparable gain. The realizable part of the $k = 10$ code carries almost all of its kNN accuracy (0.91 vs. 0.92), while $Z_\perp$ alone reaches 0.32; the remainder mostly serves reconstruction, not class structure. **(v)** Increasing k from 2 to 10 improves every quality metric of the linear methods while leaving ρ at zero; dimension and geometric mechanism are separate axes. **(vi) Output-space checks.** For the nonlinear autoencoders, 2%–6% of the reconstruction residual lies in the decoder's tangent plane at $k = 2$ and 11–12% at $k = 10$ (Table 6), so the decoded surface is close to, but not exactly at, first-order optimality at the larger width. The decoder tangent plane is somewhat better contained in the data's local 20-dimensional tangent than PCA's global plane is (47°–75° versus 58°–81° largest angle), but neither is contained: in raw pixel space the local tangent estimated from 60 neighbors is itself noisy, and this check is only weakly informative there. The clean statements remain the algebraic ones of Proposition 4.

Table 6: Checks of §5. Linear AE: departure of $W_d W_e$ from idempotence and symmetry (max entry), and the largest principal angle between $\mathrm{col}(W_d)$ and the PCA-k subspace; $\kappa(A)$ of the encoder's coordinate map. Nonlinear AE: fraction of the reconstruction residual lying in the decoder's tangent plane (first-order optimality predicts ≈ 0), and the mean over 200 points of the largest principal angle between the decoder tangent plane and the 20-dimensional local PCA tangent of the data (with PCA's global subspace as reference).

Data	k	Linear AE: $W_d W_e$			Nonlinear AE	
		idempotence	symmetry	$\angle(\mathrm{col}\, W_d,\ \mathrm{PCA})$ / $\kappa(A)$	resid. in tangent	$\angle$(dec. tangent, local) / PCA ref.
MNIST	2	5.8e-05	0.021	2.7° / 1.02	0.016	58° / 58°
	10	2.8e-04	0.056	85.5° / 1.28	0.121	69° / 75°
Fashion-MNIST	2	7.6e-05	0.017	1.1° / 1.07	0.057	47° / 58°
	10	2.6e-04	0.038	2.1° / 1.59	0.111	75° / 81°

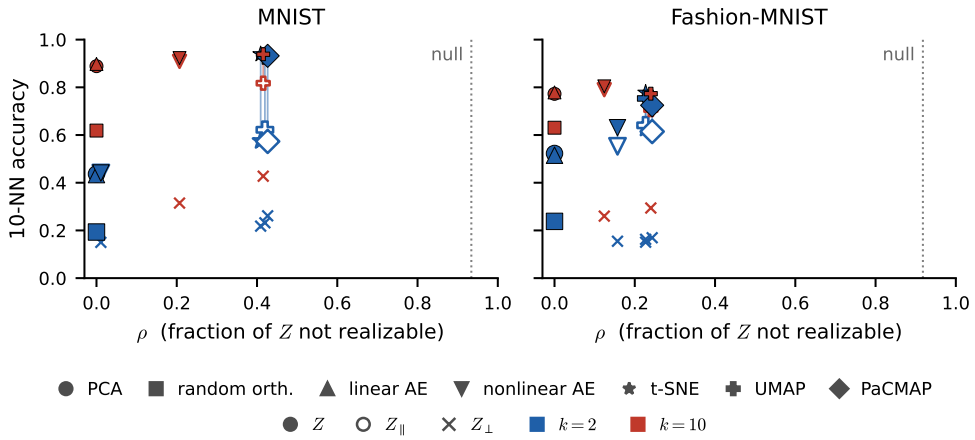

Figure 3: Where the value lives. For each method and data set, 10-NN accuracy of the full representation Z (filled), of its realizable part $Z_{\parallel}$ (hollow), and of the remainder $Z_{\perp}$ (cross), against ρ. For linear methods $Z = Z_{\parallel}$. For t-SNE, UMAP and PaCMAP the realizable part falls most of the way from the full layout toward PCA in kNN accuracy (and all the way to PCA in trustworthiness, Table 2); at $k = 2$ the remainder alone is near chance.

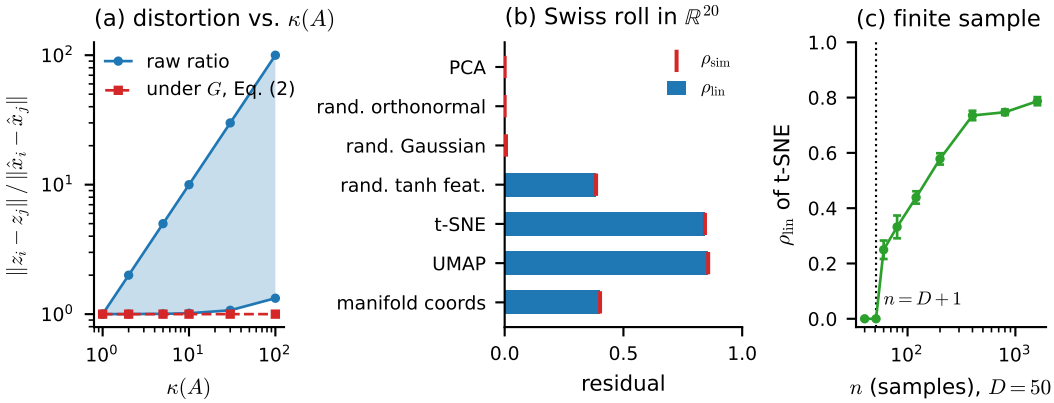

Figure 4: Synthetic tier. **(a)** Same projection, reported through $z = AU^{\top}x$ with increasing $\kappa(A)$: raw Euclidean distance ratios spread linearly with $\kappa(A)$; distances under $G = A^{-\top}A^{-1}$ match the projection exactly. **(b)** Residuals on a Swiss roll in $\mathbb{R}^{20}$. **(c)** ρ_{lin} of t-SNE on pure noise versus n at $D = 50$: identically zero for $n \le D + 1$.

E Exact Synthetic Geometry

With $D = 10$, $k = 3$, $n = 500$ Gaussian points and a generic W, the QR construction of Theorem 2 gives $\max|W - AU^{\top}| = 4.4 \times 10^{-16}$ and $\max|\hat{x} - UA^{-1}z| = 8.9 \times 10^{-16}$; Theorem 1 and Eq. (2) hold to 1.8×10^{-15} over 2000 random pairs, while the raw Euclidean ratio $\|z_i - z_j\| \,/\, \|\hat{x}_i - \hat{x}_j\|$ ranges over $[2.24, 4.30]$ for $\kappa(A) = 1.92$. Sweeping $\kappa(A)$ from 1 to 100 (Figure 4a), the raw ratio range grows to $[1.33, 99.9]$ and the Spearman correlation between reported and projected distances falls from 1.00 to 0.53, while metric-corrected distances agree to 10^{-12}. On a Swiss roll embedded in $\mathbb{R}^{20}$ (Figure 4b, Table 7), PCA and random linear maps have $\rho_{\text{lin}} \le 2 \times 10^{-15}$ as they must; two principal components of random $\tanh$ features have $\rho_{\text{lin}} = 0.385$; t-SNE and UMAP have $\rho_{\text{lin}} = 0.84$ and 0.86 with ρ_{sim} within 10^{-3}; and the true manifold coordinates have $\rho_{\text{lin}} = 0.40$: the correct answer for a curved manifold is itself not globally realizable, which motivates the local extension of §I. Figure 4(c) shows Proposition 3: t-SNE of pure noise at $D = 50$ has $\rho_{\text{lin}} \le 1.5 \times 10^{-7}$ for $n \le 51$, 0.25 at $n = 60$, and 0.79 at $n = 1600$.

Table 7: Swiss roll in $\mathbb{R}^{20}$ ($n = 2000$, $k = 2$): residuals and Spearman correlation of pairwise distances with the input.

Map	ρ_{lin}	ρ_{sim}	Spearman
PCA	$< 10^{-14}$	$< 10^{-14}$	0.85
Random orthonormal	$< 10^{-14}$	0.000	0.50
Random Gaussian	$< 10^{-14}$	0.006	0.71
Random tanh features	0.385	0.385	0.56
t-SNE	0.842	0.843	0.37
UMAP	0.856	0.857	0.30
Manifold coords (t, height)	0.400	0.402	0.44

F Richer Sources and Seed Variability

The claim that a neighborhood embedding's value lives in $Z_\perp$ is relative to the source: $Z_\perp$ is what no *linear* function of pixels can produce. Relative realizability (§4) asks the same question of a richer source $\phi(X)$, and if a modest nonlinear feature map absorbed $Z_\perp$ entirely the claim would be an artifact of the linear class. Table 8 replaces pixels by random ReLU features and by RBF Nyström maps, each at $p = 784$ features (matched to the pixels) and $p = 2000$, and adds a random *linear* expansion of the pixels to $p = 2000$, which has the same column space as X but as many regression features as the largest nonlinear sources. Because such sources have real fitting capacity at $n = 5000$ (the in-sample null falls from 0.93 to 0.77–0.79), the realizable part is evaluated *out of sample*: the map from $\phi(X)$ to Z is a ridge regression fitted on 3500 points with λ chosen on a validation split, and ρ_{out} and the utilities are measured on the remaining 1500; with this protocol the null is 1.00 for every source (slightly above 1 when the ridge map overfits). The layouts and images are exactly those of Table 2, so ρ_{in} for pixels reproduces its residuals.

Three things happen. First, capacity alone does nothing: the linear expansion to 2000 features leaves ρ_{out} at the pixel level or slightly above (0.48 vs. 0.47 for t-SNE on MNIST, 0.28 vs. 0.27 on Fashion-MNIST) and does not improve the realizable part's kNN accuracy. Second, nonlinearity does absorb part of $Z_\perp$, and most of the effect is already present at matched dimension: for t-SNE on MNIST ρ_{out} falls from 0.47 (pixels) to 0.42 and 0.37 (ReLU, $p = 784$ and 2000) and to 0.31 and 0.28 (RBF), and the held-out 10-NN accuracy of the realizable part rises from 0.52 to 0.70 with the 784-feature RBF map and 0.74 with 2000 (0.80 for UMAP); on Fashion-MNIST from 0.27 to 0.18 and 0.17 in ρ_{out} and from 0.60 to 0.68 and 0.66 in kNN. Roughly half of the gap between $Z_\parallel$ and Z is thus nonlinear structure that a generic kernel of pixel space already contains. Third, the other half is not: with the richest source the realizable part still trails the full layout by 0.19 (t-SNE, MNIST) and 0.07 (Fashion-MNIST) in held-out accuracy (the last row of each block in Table 8), at ρ_{out} far below its null. The statement is therefore source-relative: relative to pixels the neighborhood fidelity of t-SNE is almost entirely non-realizable; relative to a generic kernel source about half of its advantage becomes a projection; and a residual that survives generic kernels is what the layout-specific objective adds. Pretrained network features as source are the natural next rung. Supervised features of a small MLP trained on the same labels (rank ≈ 110) absorb less than the kernels do in-sample ($\rho = 0.35$ vs. 0.18 on MNIST), which is the expected consequence of their low rank rather than evidence about their content.

Seed variability. Table 9 repeats the $k = 2$ decomposition for t-SNE, UMAP, and the $\tanh$ autoencoder on three independent draws of 5000 images with matching method seeds. Standard deviations of ρ are ≤ 0.01 for the embeddings and 0.02 for the autoencoder on Fashion-MNIST (where different runs reach ρ between 0.16 and 0.21), and the gaps between Z and $Z_\parallel$ vary by less than 0.03; the findings of §6.1 are not seed effects.

G The Autoencoder Sweep in Detail

The single $\tanh$ autoencoder above is one point; Table 10 (Appendix D) and Figure 2 sweep encoder depth (1, 2, 3 hidden layers per side, widths 256; $512, 256$; $512, 256, 128$), activation ($\tanh$, ReLU at depth 2), and $k \in \{2, 10, 32\}$, training on 4000 images and reporting held-out reconstruction on the

Table 8: Relative realizability with richer sources ($k = 2$; the images and layouts of Table 2). p: number of features; rank in parentheses. ρ_{in}: in-sample residual (Proposition 1). ρ_{out} and kNN $Z_{\|}$: ridge map $\phi(X) \to Z$ fitted on 3500 points (λ by validation), evaluated on the 1500 held-out points. The last row of each block is the held-out kNN accuracy of the full layout, the ceiling for kNN $Z_{\|}$.

		t-SNE		UMAP		null	
Data	Source $\phi(X)$ (rank)	ρ_{in} / ρ_{out}	kNN $Z_{\|}$	ρ_{in} / ρ_{out}	kNN $Z_{\|}$	ρ_{in}	ρ_{out}
MNIST	pixels X (642)	0.41 / 0.47	0.52	0.42 / 0.48	0.56	0.93	1.00
	PCA-50 of X (50)	0.47 / 0.48	0.52	0.49 / 0.50	0.53	1.00	1.00
	linear expansion, $p = 2000$ (642)	0.41 / 0.48	0.50	0.42 / 0.49	0.52	0.93	1.05
	random ReLU, $p = 784$ (733)	0.34 / 0.42	0.58	0.35 / 0.42	0.63	0.92	1.00
	random ReLU, $p = 2000$ (1883)	0.23 / 0.37	0.62	0.24 / 0.37	0.66	0.79	1.00
	RBF Nyström, $p = 784$ (784)	0.26 / 0.31	0.70	0.26 / 0.30	0.75	0.92	1.00
	RBF Nyström, $p = 2000$ (2000)	0.18 / 0.28	0.74	0.18 / 0.27	0.80	0.77	1.00
	supervised MLP, $p = 128$ (111)	0.35 / 0.37	0.67	0.35 / 0.37	0.72	0.99	1.00
	full layout Z (held out)	—	0.94	—	0.93		
Fashion-MNIST	pixels X (784)	0.23 / 0.27	0.60	0.23 / 0.26	0.61	0.91	1.00
	PCA-50 of X (50)	0.27 / 0.28	0.59	0.27 / 0.27	0.62	0.99	1.00
	linear expansion, $p = 2000$ (784)	0.23 / 0.28	0.59	0.23 / 0.27	0.60	0.91	1.06
	random ReLU, $p = 784$ (746)	0.18 / 0.23	0.64	0.16 / 0.20	0.65	0.92	1.00
	random ReLU, $p = 2000$ (1920)	0.13 / 0.21	0.64	0.10 / 0.18	0.66	0.78	1.00
	RBF Nyström, $p = 784$ (784)	0.15 / 0.17	0.68	0.13 / 0.16	0.64	0.92	1.00
	RBF Nyström, $p = 2000$ (2000)	0.10 / 0.17	0.66	0.09 / 0.14	0.65	0.77	1.00
	supervised MLP, $p = 128$ (117)	0.27 / 0.28	0.65	0.23 / 0.25	0.63	0.99	1.00
	full layout Z (held out)	—	0.73	—	0.72		

Table 9: Mean $\pm$ s.d. over three independent subsamples and seeds ($k = 2$).

Data	Method	ρ	Trust. Z	Trust. $Z_{\|}$	kNN Z	kNN $Z_{\|}$
MNIST	t-SNE	0.407 ± 0.001	0.981 ± 0.001	0.767 ± 0.002	0.935 ± 0.003	0.584 ± 0.008
	UMAP	0.418 ± 0.002	0.959 ± 0.002	0.754 ± 0.002	0.927 ± 0.007	0.619 ± 0.004
	Nonlinear autoencoder	0.010 ± 0.001	0.745 ± 0.004	0.744 ± 0.004	0.440 ± 0.001	0.438 ± 0.003
Fashion-MNIST	t-SNE	0.224 ± 0.007	0.988 ± 0.000	0.919 ± 0.001	0.769 ± 0.017	0.641 ± 0.003
	UMAP	0.232 ± 0.009	0.980 ± 0.001	0.896 ± 0.010	0.755 ± 0.011	0.615 ± 0.027
	Nonlinear autoencoder	0.189 ± 0.021	0.952 ± 0.002	0.914 ± 0.003	0.640 ± 0.022	0.570 ± 0.012

remaining 1000 alongside the certificates of §5. Three regularities hold across the 24 configurations. **(a) In end-to-end training the gain over PCA accompanies $Z_{\perp}$; it does not require it.** The only configuration with $\rho \approx 0$ (MNIST, $k = 2$, one $\tanh$ layer, $\rho = 0.015$) has exactly PCA's held-out error, and every configuration with $\rho > 0.05$ beats PCA. At $k = 10$ and $k = 32$ the $\tanh$ gain increases with depth together with ρ on both data sets (MNIST, $k = 10$: $\rho = 0.16, 0.22, 0.24$ against gain $0.068, 0.090, 0.105$; $k = 32$: $\rho = 0.08, 0.10, 0.13$ against $0.037, 0.047, 0.055$). At $k = 2$ the two decouple: on Fashion-MNIST the three-layer model has the larger ρ (0.31 vs. 0.17) and the smaller gain (0.078 vs. 0.104), and on MNIST the three-layer model has the smaller ρ and the larger gain. ReLU autoencoders reach the largest ρ in five of the six settings; on MNIST they overfit (train error 0.32 vs. test 0.46 at $k = 32$) and their held-out gain is below $\tanh$'s at $k \geq 10$, while on Fashion-MNIST they give the largest held-out gain of the sweep at $k = 2$ (0.141) and $k = 10$ (0.087). Proposition 2 does not say that a nonlinear decoder needs $Z_{\perp}$, and Table 4 tests whether it does. A nonlinear decoder of the same architecture trained on the PCA scores, a code with $Z_{\perp} = 0$ by construction, beats PCA in all six settings (gain 0.023–0.095), and matches or exceeds an end-to-end autoencoder at $k = 10$ and $k = 32$. What end-to-end training adds is at the tightest bottleneck: at $k = 2$ the autoencoder's own code gains more than the PCA scores do (0.065 vs. 0.025 on MNIST, 0.094 vs. 0.064 on Fashion-MNIST), and the code it chooses has $Z_{\perp} \neq 0$. (The same one-layer architecture trained in JAX reaches $\rho = 0.31$ at $k = 2$ on MNIST, where the scikit-learn run above stayed near-linear: which regime training lands in is a property of the optimization, not of the architecture.) Its decoder then depends on that $Z_{\perp}$: fed only the realizable part of its code, the same decoder falls below PCA on MNIST at $k = 2$ (-0.073) and $k = 32$ (-0.029); on Fashion-MNIST it keeps about half of its gain at $k = 2$ and nearly all of it at $k = 10$ and 32. Fed only $Z_{\perp}$, placed at the code's mean, it does worse than predicting the mean image (relative error 1.01–1.25): the remainder is not a separate channel of information but a correction that is useful only on top of the projection. So ρ measures how far a trained code departs from a projection, and the gain over PCA is a joint property of the decoder's nonlinearity and the code it is given; $Z_{\perp} \neq 0$ is neither necessary nor sufficient for

it. **(b) No sign of a mirror image.** Across 720 Jacobians the count of eigenvalues of modulus $> \frac{1}{2}$ with negative real part is 0 for every configuration with $k \geq 10$ and at most 0.07 per point at $k = 2$ (two of 30 points in the worst case; Table 10 rounds this to 0.1). By Proposition 6(iii) this excludes a reflection only at the points it would fix; the decisive evidence against a mirrored reconstruction is Proposition 5 together with held-out errors below PCA's, since a reconstruction $g \circ R$ with g a non-trivial isometry incurs the error $\|x - g(x)\|$ at every point g moves. **(c) Recovering every point is not the same as recovering the space.** One-layer $\tanh$ autoencoders are, to good approximation, projections: $J = d(\mathcal{D} \circ E)$ has 1.9, 5.2, and 25.6 eigenvalues near 1 for $k = 2, 10, 32$ on MNIST, trace 1.9, 9.3, 31.9, and idempotence defect 0.07–0.38. Deeper autoencoders reconstruct better while ceasing to be projections: at three layers the defect is 0.28–0.76, the near-unit eigenvalue count falls to 0.4, 2.7, and 17.7, and the map-level cycle defect $\|R(R(x)) - R(x)\| \,/\, \|R(x)\|$ on held-out data is 0.10–0.20 (and 0.65 for the ReLU $k = 2$ model on MNIST). These networks send each training point back to itself but do not fix their own decoded manifold pointwise ($E \circ \mathcal{D} \neq \mathrm{id}$), so a point on $M_{\mathcal{D}}$ that is not a training point is moved by R. The distinction between the two guarantees of §5, identity on the data versus projector structure, is therefore not academic: standard training delivers the first and, beyond one hidden layer, not the second. Strikingly, the idempotence defect is close to a single increasing function of ρ across all data sets, depths, widths, and activations (Figure 2, right): how far an autoencoder is from a projection of its input predicts how far its reconstruction map is from a projection onto its own output, which suggests the two residuals are one quantity seen from the two ends of the network.

Table 10: Nonlinear autoencoder sweep ($n_{\mathrm{train}} = 4000$, $n_{\mathrm{test}} = 1000$). ρ: residual of the bottleneck code w.r.t. the input. Recon. (test): relative held-out reconstruction error, with PCA at the same k for reference; gain = PCA − AE. Certificates of Proposition 6 at 30 data points: idempotence defect $\|J^2 - J\| \,/\, \|J\|$ of $J = d(\mathcal{D} \circ E)$, number of eigenvalues within 0.2 of 1 (predicted k), number of eigenvalues of modulus $> \frac{1}{2}$ with negative real part (a reflection would give ≥ 1), and the map-level cycle defect $\|R(R(x)) - R(x)\| \,/\, \|R(x)\|$ on held-out data.

Data	AE (act., depth)	k	ρ	kNN	Recon. test	PCA test	gain	idemp. defect	#eig≈ 1	#neg eig	cycle defect
MNIST	tanh, 1	2	0.015	0.427	0.912	0.912	0.000	0.07	1.9	0.0	0.087
	tanh, 2	2	0.389	0.609	0.853	0.912	+0.059	0.53	0.6	0.1	0.175
	tanh, 3	2	0.374	0.682	0.825	0.912	+0.086	0.74	0.4	0.1	0.131
	relu, 2	2	0.397	0.755	0.823	0.912	+0.088	0.59	0.5	0.1	0.651
	tanh, 1	10	0.158	0.897	0.645	0.713	+0.068	0.32	5.2	0.0	0.144
	tanh, 2	10	0.215	0.920	0.622	0.713	+0.090	0.37	3.2	0.0	0.171
	tanh, 3	10	0.239	0.925	0.608	0.713	+0.105	0.43	2.7	0.0	0.192
	relu, 2	10	0.261	0.909	0.637	0.713	+0.076	0.54	1.6	0.0	0.250
	tanh, 1	32	0.081	0.930	0.465	0.502	+0.037	0.19	25.6	0.0	0.094
	tanh, 2	32	0.101	0.931	0.455	0.502	+0.047	0.21	24.6	0.0	0.116
	tanh, 3	32	0.133	0.936	0.446	0.502	+0.055	0.28	17.7	0.0	0.147
	relu, 2	32	0.226	0.921	0.464	0.502	+0.038	0.49	5.0	0.0	0.185
Fashion-MNIST	tanh, 1	2	0.118	0.617	0.659	0.730	+0.071	0.38	0.8	0.0	0.109
	tanh, 2	2	0.171	0.637	0.626	0.730	+0.104	0.37	0.8	0.0	0.109
	tanh, 3	2	0.308	0.642	0.652	0.730	+0.078	0.76	0.5	0.0	0.196
	relu, 2	2	0.279	0.677	0.589	0.730	+0.141	0.46	0.5	0.0	0.179
	tanh, 1	10	0.101	0.776	0.488	0.531	+0.043	0.26	6.0	0.0	0.085
	tanh, 2	10	0.124	0.782	0.474	0.531	+0.057	0.31	4.6	0.0	0.105
	tanh, 3	10	0.164	0.792	0.465	0.531	+0.066	0.36	3.2	0.0	0.117
	relu, 2	10	0.231	0.787	0.444	0.531	+0.087	0.37	2.7	0.0	0.112
	tanh, 1	32	0.061	0.801	0.399	0.422	+0.023	0.18	26.3	0.0	0.068
	tanh, 2	32	0.090	0.805	0.392	0.422	+0.030	0.22	23.2	0.0	0.089
	tanh, 3	32	0.116	0.797	0.388	0.422	+0.034	0.28	18.3	0.0	0.102
	relu, 2	32	0.221	0.804	0.400	0.422	+0.022	0.43	5.5	0.0	0.122

H PaCMAP, Regularized Autoencoders and CIFAR-10 in Detail

PaCMAP. PaCMAP adds mid-near pairs to the attractive and repulsive terms precisely to keep global structure that t-SNE and UMAP discard (Wang et al., 2021). Its layouts split the same way (Table 2): $\rho = 0.43$ on MNIST and 0.24 on Fashion-MNIST, slightly above t-SNE's and UMAP's residuals, with a realizable part at PCA's trustworthiness (0.745 vs. 0.747 on MNIST, 0.894 vs. 0.914 on Fashion-MNIST) and well below the full layout's kNN accuracy (0.57 vs. 0.93; 0.61 vs. 0.72). Preserving more global geometry does not make the layout more of a projection of pixel space.

Table 11: Regularized autoencoders on the architecture of Table 4: residual ρ of the code (the mean μ for VAEs), held-out reconstruction gain over PCA, and 10-NN accuracy of the code. Contractive: penalty $\lambda \|J_E(x)\|_F^2$. VAE: loss $\|x - \mathcal{D}(z)\|^2 + \beta\,\mathrm{KL}(q(z \mid x) \,\|\, \mathcal{N}(0, I))$.

		$k = 2$			$k = 10$		
Data	Model	ρ	gain	kNN	ρ	gain	kNN
MNIST	plain AE	0.313	+0.065	0.625	0.157	+0.097	0.920
	contractive, $\lambda = 1$	0.301	+0.069	0.617	0.096	+0.062	0.900
	contractive, $\lambda = 10$	0.262	+0.063	0.619	0.007	−0.001	0.890
	VAE, $\beta = 1$	0.299	+0.076	0.658	0.096	+0.085	0.911
	β-VAE, $\beta = 4$	0.189	+0.044	0.568	0.062	−0.124	0.804
Fashion-MNIST	plain AE	0.163	+0.094	0.636	0.098	+0.058	0.794
	contractive, $\lambda = 1$	0.163	+0.098	0.638	0.068	+0.048	0.795
	contractive, $\lambda = 10$	0.158	+0.099	0.623	0.008	−0.001	0.778
	VAE, $\beta = 1$	0.174	+0.120	0.653	0.096	+0.038	0.779
	β-VAE, $\beta = 4$	0.151	+0.097	0.629	0.113	−0.059	0.723

Regularized and probabilistic autoencoders. Table 11 trains contractive autoencoders (Rifai et al., 2011), which penalize $\lambda \|J_E(x)\|_F^2$, and variational autoencoders with KL weight β (Kingma & Welling, 2014; Higgins et al., 2017) on the architecture of Table 4. Regularization suppresses $Z_\perp$, and how much depends on its strength and on k. At $k = 10$ a strong contraction ($\lambda = 10$) drives ρ from 0.16 to 0.007 on MNIST and from 0.10 to 0.008 on Fashion-MNIST, and the gain over PCA falls to zero with it: the penalty makes the encoder an affine projection and the autoencoder reverts to PCA. The β-VAE at $\beta = 4$ lowers ρ on MNIST ($0.31 \to 0.19$ at $k = 2$, $0.16 \to 0.06$ at $k = 10$) and pays in reconstruction, falling below PCA at $k = 10$ on both data sets. At $k = 2$ the contractive penalty barely moves ρ on Fashion-MNIST and lowers it modestly on MNIST ($0.31 \to 0.26$ at $\lambda = 10$) at no cost in gain, and the unit-weight VAE has the largest gain at $k = 2$ on both data sets. Within end-to-end training, then, the regularizers that make the latent map smoother make it more realizable, and where they make it realizable they remove the gain over PCA; the decoder-on-PCA control of Table 4 shows this is a property of what joint training finds, not of the decoder's capacity.

Natural images: CIFAR-10. Table 3 repeats the decomposition on the $10{,}000$ CIFAR-10 test images (Krizhevsky, 2009), $D = 3072$ RGB pixels, natural images on which pixel-space PCA captures far less class structure than on digits (10-NN accuracy 0.17 at $k = 2$). The null residual is 0.83 ($\operatorname{rank} X = 3072$, $n/D = 3.3$). The layouts are *more* realizable than on digits: $\rho = 0.30$ (t-SNE), 0.25 (PaCMAP) and 0.23 (UMAP), so a fixed projection of pixel space carries 91–95% of their energy. Yet that part is exactly PCA-grade on both utilities: trustworthiness 0.80–0.81 against PCA's 0.82, and 10-NN accuracy 0.165–0.174 against PCA's 0.173, with realized subspaces only $13°$–$18°$ from PCA's. Everything t-SNE adds over PCA on CIFAR-10 (trustworthiness 0.90, 10-NN 0.29) is carried by $Z_\perp$. Energy dominance of the realizable part thus survives, and strengthens, on natural images, while the separation between energy and value becomes sharper, not weaker.

I Local Realizability

Global realizability asks for one linear view of the source that explains the whole representation, and for curved data that is the wrong question: the true coordinates of a Swiss roll are a smooth, invertible function of the source yet have $\rho_{\mathrm{lin}} = 0.40$ (§E). The decomposition localizes directly.

Definition 2 (Local realizability)**.** For $x_i \in X$ let $N_K(i)$ be its K nearest neighbors in the source, let $T_i \in \mathbb{R}^{K \times d}$ be the d leading local principal coordinates of $\{x_j : j \in N_K(i)\}$ after centering, and let Z_i be the corresponding rows of Z, centered. The local residual at x_i is $\rho_{\mathrm{loc}}(i; K, d) = \left\| Z_i - \Pi_{[T_i, \mathbf{1}]} Z_i \right\|_F / \|Z_i\|_F$, and $\rho_{\mathrm{loc}}(K, d)$ is its median over x_i.

Proposition 7 (Local realizability of smooth maps)**.** *Let the data lie on a C^2 d-manifold M and let $f = \phi \circ \pi$ with π a chart of M and $\phi \in C^2$, and suppose f is non-degenerate on M: df_x restricted to $T_x M$ is nonzero, with $\|df_x|_{T_x M}\|$ bounded below on the region considered. Then $\rho_{\mathrm{loc}}(K, d) \to 0$ as the neighborhood radius $r_K \to 0$, at rate $O(r_K)$ set by the second derivatives of ϕ, the curvature*

of M and that lower bound; any representation with ρ_{loc} bounded away from zero at all scales is not a non-degenerate C^2 function of the source. (At a critical point of f both the residual and $\|Z_i\|_F$ are $O(r_K^2)$ and the ratio need not vanish.) Conversely, $\rho_{\mathrm{loc}} \equiv 0$ trivially whenever $K \leq d+1$, so K must exceed $d+1$ by a margin (the local form of Proposition 3), and the local PCA truncation to d dimensions means that a globally realizable map can have $\rho_{\mathrm{loc}} > 0$ if its realized subspace is not contained in the local tangent: PCA is the reference for that effect.

Proof. Within a neighborhood of radius r_K, M is within $O(r_K^2)$ of its tangent plane and f is within $O(r_K^2)$ of its affine Taylor approximation in tangent coordinates, so the affine fit on T_i leaves a residual of norm $O(r_K^2)$, while non-degeneracy gives $\|Z_i\|_F = \Theta(r_K)$; the ratio is $O(r_K)$. For the triviality: T_i and Z_i are centered, so all their columns lie in $\mathbf{1}^\perp \subset \mathbb{R}^K$, of dimension $K-1$, and the intercept column contributes nothing to the projection of a centered Z_i. The d columns of T_i generically span $\mathbf{1}^\perp$ iff $d \geq K-1$, which is Proposition 3 with D replaced by d. □

Local realizability is therefore not a weaker form of the global property but a different attribute: it measures the *scale* at which a representation stops being a projection of its source, and it separates "curved but smooth" from "not a function of the source at all". Table 12 reports $\rho_{\mathrm{loc}}(K)$ across neighborhood sizes. On the Swiss roll, the manifold coordinates have $\rho_{\mathrm{loc}} = 0.05$ at $K = 10$, 0.02 at $K = 80$, then 0.11, 0.20, 0.32 at $K = 160, 400, 1000$, approaching the global 0.40: curved but smooth, exactly as Proposition 7 predicts, with the rise marking the scale at which the roll's curvature is felt. t-SNE and UMAP show the same shape (0.10–0.27 up to $K = 160$, then 0.41–0.42 at $K = 400$ and 0.69–0.71 at $K = 1000$), while random $\tanh$ features sit at 0.39–0.46 at every scale, and the null at 0.82–1.00: the local residual separates a smooth reparameterization of the manifold (t-SNE, UMAP, the true coordinates) from a map that is not a function of the local tangent at any scale. On the image data the picture is different and is the informative one. PCA, globally realizable, has $\rho_{\mathrm{loc}} = 0.09$–$0.28$ because a 10-dimensional local tangent does not contain the global principal directions (the truncation effect of Proposition 7); that is the reference level. Against it, the nonlinear autoencoder is indistinguishable from PCA on MNIST and 0.06–0.15 above it on Fashion-MNIST, while t-SNE and UMAP are 0.07–0.36 above PCA at every K on both data sets and, unlike on the Swiss roll, never fall to PCA's level as K shrinks. Neighborhood embeddings of image data are not locally realizable at any of the scales tested: they are not smooth functions of pixel-space neighborhoods, which is consistent with their lookup-table parameterization and with the tearing and cluster-separation they are known to perform. This is the local counterpart of the global finding in §6.1, and it is the sense in which the "curved manifold" objection to global realizability does not rescue t-SNE on images: the remainder $Z_\perp$ there is not curvature.

J Depth Without Dimension Change, and Layer Linearization

For a four-hidden-layer ReLU MLP of constant width 128 trained on each data set we compute $\rho(H_{\ell-1}, H_\ell)$ (how much of layer ℓ is a fixed projection of layer $\ell-1$), $\rho(X, H_\ell)$ (of the input), and the participation ratio $\mathrm{PR}(H_\ell) = (\sum \lambda_i)^2 / \sum \lambda_i^2$ of each layer's covariance as a global dimension estimate (with $n = 5000 \gg 128$ its finite-sample bias is small; a corrected estimator is given by Chun et al. (2026)). Table 13 shows the profile. The first layer maps 784 pixels to 128 units and has $\rho(H_0, H_1) = 0.22$–0.25. The three transitions between layers of *identical width* ($128 \to 128$) are not projections of one another either: $\rho(H_{\ell-1}, H_\ell)$ is 0.14–0.19 at the second layer and decreases to 0.09–0.10 by the fourth (later ReLU layers act nearly linearly on an already-shaped representation), while $\rho(X, H_\ell)$ rises monotonically to 0.28–0.33. The participation ratio moves on its own schedule: on MNIST it drops from 29 at the input to 10 after one layer and 6.6 at depth four, while on Fashion-MNIST it *rises* from 7.9 to 9.3 at the first layer before falling to 6.6, even though the residual profiles of the two networks are nearly identical. The network changes geometry at every layer while changing dimension at none, and residual and dimension are different readings of the same layer. This is the setting in which "dimensionality reduction" is not even nominally the right description, and where ρ, dimension estimates (Chun et al., 2026), and topological summaries (Moisescu-Pareja et al., 2026) each capture a distinct attribute.

Table 12: Local realizability $\rho_{\mathrm{loc}}(K, d)$ (median over 300 anchors) as a function of neighborhood size K, with the global ρ_{lin} for reference. Swiss roll in $\mathbb{R}^{20}$: $d = 3$. MNIST and Fashion-MNIST ($k = 2$ embeddings of §6.1): $d = 10$. The null is a random Z.

Swiss roll in $\mathbb{R}^{20}$ ($d = 3$) Method	$K = 10$	$K = 20$	$K = 40$	$K = 80$	$K = 160$	$K = 400$	$K = 1000$	global ρ_{lin}
PCA	0.05	0.02	0.01	0.00	0.00	0.00	0.00	$< 10^{-14}$
Random tanh features	0.39	0.42	0.41	0.42	0.45	0.46	0.43	0.385
t-SNE	0.10	0.16	0.20	0.20	0.22	0.42	0.71	0.842
UMAP	0.22	0.26	0.27	0.25	0.25	0.41	0.69	0.856
Manifold coords	0.05	0.04	0.03	0.02	0.11	0.20	0.32	0.400
null (random Z)	0.82	0.92	0.96	0.98	0.99	1.00	1.00	0.996
MNIST ($d = 10$) Method	$K = 30$	$K = 60$	$K = 120$	$K = 250$	$K = 500$	$K = 1000$		global ρ_{lin}
PCA	0.24	0.27	0.28	0.26	0.23	0.19		$< 10^{-14}$
Nonlinear autoencoder	0.24	0.28	0.28	0.26	0.23	0.19		0.011
t-SNE	0.37	0.47	0.52	0.54	0.54	0.52		0.409
UMAP	0.31	0.43	0.50	0.55	0.56	0.55		0.420
null (random Z)	0.81	0.91	0.96	0.98	0.99	1.00		0.934
Fashion-MNIST ($d = 10$) Method	$K = 30$	$K = 60$	$K = 120$	$K = 250$	$K = 500$	$K = 1000$		global ρ_{lin}
PCA	0.15	0.15	0.15	0.13	0.12	0.09		$< 10^{-14}$
Nonlinear autoencoder	0.21	0.23	0.24	0.24	0.24	0.24		0.157
t-SNE	0.35	0.42	0.45	0.46	0.45	0.40		0.227
UMAP	0.29	0.35	0.37	0.38	0.41	0.39		0.227
null (random Z)	0.82	0.91	0.96	0.98	0.99	0.99		0.918

Table 13: Depth profile of a width-128 ReLU MLP ($n = 5000$): residual from the previous layer, residual from the input, and participation ratio, per layer. The network is trained on 4000 of the images; its end-to-end test accuracy on the other 1000 is 0.936 (MNIST) and 0.830 (Fashion-MNIST), and it is the network linearized in Table 14.

Data	layer ℓ	width	$\rho(H_{\ell-1}, H_\ell)$	$\rho(X, H_\ell)$	PR(H_ℓ)
MNIST	0	784	—	—	29.4
	1	128	0.215	0.215	10.4
	2	128	0.139	0.233	7.7
	3	128	0.098	0.252	6.7
	4	128	0.086	0.281	6.6
Fashion-MNIST	0	784	—	—	7.9
	1	128	0.252	0.252	9.3
	2	128	0.190	0.287	7.7
	3	128	0.130	0.310	7.1
	4	128	0.096	0.334	6.6

From a diagnostic to a prediction: the linearization bound

A residual is a description; a theory has to predict. The natural prediction is that a layer with small ρ can be *replaced* by its realizable part at bounded cost.

Proposition 8 (Layer linearization)**.** *Let $F_{>\ell}$ denote the map from layer ℓ's activations to the logits, with Lipschitz constant $L_{>\ell} \le \prod_{j>\ell} \|W_j\|_2$ for a ReLU network. Replace h_ℓ by its realizable part $\tilde{h}_\ell$, the least-squares affine map of $h_{\ell-1}$, so that $\sum_i \left\|h_{\ell,i} - \tilde{h}_{\ell,i}\right\|^2 = \rho^2 \|H_\ell\|_F^2$ with $\rho = \rho(H_{\ell-1}, H_\ell)$. Then for every sample $\left\|F_{>\ell}(h_\ell) - F_{>\ell}(\tilde{h}_\ell)\right\|_2 \le L_{>\ell} \left\|h_\ell - \tilde{h}_\ell\right\|_2$, and the predicted class is unchanged whenever*

$$L_{>\ell} \left\|h_\ell - \tilde{h}_\ell\right\|_2 < \frac{m(x)}{\sqrt{2}}, \qquad m(x) = \textit{logit margin of } x. \tag{6}$$

In particular the accuracy drop is at most the fraction of test points violating (6)*, and by Markov's inequality at most $2L_{>\ell}^2 \rho^2 \|H_\ell\|_F^2 / (n\, m_{\min}^2)$ when all margins exceed $m_{\min}$.*

Table 14: Layer linearization in the width-128 ReLU MLP of Table 13. Each hidden layer is replaced, one at a time, by its ridge-regularized affine map from the previous layer (fitted on the training split); the remaining layers run unchanged. ρ_{test} is the residual of that map on the test split; $L_{>\ell}$ is the spectral-norm Lipschitz bound of the layers above; "uncertified" is the fraction of test points for which (6) fails (an upper bound on the possible accuracy drop). A negative drop is a gain.

Data	layer ℓ	ρ_{train}	ρ_{test}	$L_{>\ell}$	acc. full	acc. linearized	drop	uncertified
MNIST	1	0.211	0.246	34.9	0.936	0.900	$+0.036$	1.00
	2	0.139	0.144	11.7	0.936	0.928	$+0.008$	0.83
	3	0.098	0.103	4.5	0.936	0.936	0.000	0.30
	4	0.085	0.091	1.8	0.936	0.934	$+0.002$	0.15
Fashion-MNIST	1	0.244	0.304	75.1	0.830	0.806	$+0.024$	1.00
	2	0.190	0.194	20.9	0.830	0.837	-0.007	0.98
	3	0.130	0.135	6.6	0.830	0.841	-0.011	0.68
	4	0.096	0.099	2.1	0.830	0.830	0.000	0.29

Proof. The first inequality is the definition of the Lipschitz constant. A change of predicted class requires the logit vector to move by at least $m(x)/\sqrt{2}$ in Euclidean norm (the difference of the top two logits must change sign, and the map $v \mapsto (v_a - v_b)/\sqrt{2}$ is 1-Lipschitz). The Markov step bounds the fraction of samples with $\left\|h_\ell - \tilde{h}_\ell\right\|^2 \geq m_{\min}^2/(2L_{>\ell}^2)$ by the ratio of the mean squared perturbation $\rho^2 \|H_\ell\|_F^2 / n$ to that threshold. □

Table 14 tests the prediction by actually linearizing each layer of the network of Table 13 (fitting a ridge-regularized affine map on the training split, with the ridge parameter chosen on a validation split, applying it on the test split, and running the remaining layers unchanged). Linearizing the first layer ($\rho_{\text{test}} = 0.25$ and 0.30) costs 3.6 and 2.4 accuracy points on MNIST and Fashion-MNIST; linearizing any later layer ($\rho_{\text{test}} = 0.09$–$0.19$) costs at most 0.8 points, and on Fashion-MNIST the linearized second and third layers even gain 0.7 and 1.1 points. The per-sample certificate (6) is reported as the fraction of test points it cannot certify; with the spectral-norm Lipschitz product it is vacuous at the first layer ($L_{>1} = 35$ and 75) and becomes informative only at depth (15% and 29% uncertified at layer 4), as such bounds are. What this experiment can and cannot say should be stated carefully. The quantity Proposition 8 controls is the composite $L_{>\ell}\left\|h_\ell - \tilde{h}_\ell\right\|$, not ρ alone. In this feed-forward network $\rho(H_{\ell-1}, H_\ell)$, $L_{>\ell}$ and the number of nonlinearities between layer ℓ and the logits all decrease with depth, so the observed ordering of accuracy drops is consistent with each of them and does not show that ρ ranks layers by itself. The claim we make is the certificate: ρ is the computable size of the replacement, the Lipschitz factor converts it into a bound on the logits, and their product, evaluated per sample against the margin, bounds the accuracy that linearization can cost. Separating the two factors requires networks in which they vary independently, such as residual architectures, which §J leaves to future work.

Relation to lifted representation geometry

The per-layer residual is the scalar face of a matrix-valued object. Writing the lifted representation of a network as $\Phi(x) = (x, h_1, \ldots, h_L)$ in the product of all layer spaces, its pullback metric is $G_\Phi = I + \sum_\ell J_\ell^\top J_\ell$ with J_ℓ the cumulative Jacobian of layer ℓ, and $C_\ell = J_\ell^\top J_\ell$ is layer ℓ's contribution. A layer whose map from $H_{\ell-1}$ is well approximated by a fixed linear map W is one with $\rho(H_{\ell-1}, H_\ell) \approx 0$, and then C_ℓ is, to first order, $W^\top W$ pulled back through the earlier layers: such a layer admits a low-rank direct map and is redundant in the lifted geometry. The depth profile of Table 13, in which $\rho(H_{\ell-1}, H_\ell)$ falls with depth, is first evidence for that redundancy hypothesis in a small network. We state it as a conjecture: *layers with small relative residual are exactly those that a depth-coarse-grained direct map can absorb without loss.* Testing it on residual networks and transformers, where the residual stream makes $\rho(H_{\ell-1}, H_\ell)$ small by construction and the informative quantity becomes the residual of the block's update, is the natural continuation and is left to future work.

K Related Work in Detail

Representation geometry at ICLR. Moisescu-Pareja et al. (2026) treat the representations of modular-addition networks as manifolds and show with persistent homology and PCA that architecturally different networks implement the same algorithm through topologically equivalent representations (discs and tori); they study the manifold, we study the map that produced it. Chun et al. (2026) ask what the dimension of a representation means, show the participation ratio is biased at finite n, and give a corrected estimator that extends to curved manifolds by local weighting; they measure how many dimensions, we measure how much of them is projection, and Proposition 3 is the same finite-sample failure in our quantity. Alshammari et al. (2025) (I-Con) unify clustering, spectral methods, dimensionality reduction, contrastive and supervised learning as integrated KL minimization between a supervisory and a learned conditional distribution; their treatment of SNE is our motivating cross-space embedding, and their objective's invariance group fixes the coordinate class in Definition 1. Uscidda et al. (2025) regularize the distortion of a map with the Gromov–Monge gap; Eq. (2) is its exact linear case, and Proposition 4 explains why distribution-level objectives need such a term while pointwise reconstruction does not. Gupta et al. (2024) design embeddings so that input transformations act as rotations of the embedding; we audit, after the fact, whether an embedding factors through a source subspace. Alleman et al. (2024) show that task structure and nonlinearity jointly shape representational geometry; our depth profile is one scalar reading of that geometry.

Classical dimensionality reduction and autoencoders. PCA and its subspace reading are standard (Jolliffe & Cadima, 2016); Krzanowski (2005) already draws the distinction between techniques whose axes are linear combinations of the variables and techniques that find a subspace directly or implicitly, which is the distinction Theorem 2 makes exact. Spectral manifold methods, Laplacian eigenmaps and diffusion maps (Belkin & Niyogi, 2003; Coifman & Lafon, 2006), produce coordinates from graph eigenvectors and are cross-space embeddings in our sense; surveys of the area (Izenman, 2012; Meilă & Zhang, 2024) emphasize that their reliability depends on neighborhood and scale choices, which is exactly the dependence the local residual of §I exposes. Neighborhood embeddings that balance local and global structure with additional attractive and repulsive terms, TriMap and PaCMAP (Amid & Warmuth, 2019; Wang et al., 2021), are cross-space embeddings of the same kind; §H decomposes PaCMAP and finds the same split as for t-SNE and UMAP. Lee & Verleysen (2011) explain the success of SNE-type methods through shift-invariant similarities that evade distance concentration, an argument about *why* the target-space objective is the right one that is complementary to our question of what the resulting layout retains of the source. Espadoto et al. (2021) benchmark projection techniques with quality metrics including the trustworthiness we use. Supervised subspace selection (Tao et al., 2009) and localized projections with a gating model (Gönen & Alpaydın, 2010) are same-space projections chosen by a discriminative criterion, globally or per region; the latter is the supervised ancestor of local realizability. Shiebler (2022) unifies manifold-learning algorithms as functors, a categorical counterpart to I-Con's information-geometric unification. Kernel PCA projects inside a feature space (Schölkopf et al., 1998), and is realizable relative to $\mathcal{H}$ but not to $\mathbb{R}^D$ (Appendix A). Linear autoencoders recover the principal subspace but not the axes (Baldi & Hornik, 1989; Plaut, 2018); Theorem 2 says why. Latent non-identifiability under reparameterization is the elementary form of results in disentanglement (Locatello et al., 2019; Bengio et al., 2013); Theorem 3 is the version the decomposition needs. Regularized and probabilistic autoencoders, contractive penalties on the encoder Jacobian (Rifai et al., 2011) and variational objectives with a weighted KL term (Kingma & Welling, 2014; Higgins et al., 2017), add exactly the symmetry-breaking terms that Theorem 3 says are needed to pin latent axes; §H measures how they move ρ.

Representation similarity. Regression- and CCA-based indices (Raghu et al., 2017; Kornblith et al., 2019) compare two representations up to a transformation class. Asymmetry is not what distinguishes ρ_{lin}: the linear-regression index analyzed by Kornblith et al. (2019) is already asymmetric and equals $1 - \rho^2_{\text{lin}}$. The difference lies in what is done with the fit: Theorem 2 identifies the fitted matrix as the coordinates of one orthogonal projection of the source, and the decomposition evaluates $Z_{\|}$ and $Z_{\perp}$ separately as representations, which a score does not. Shape metrics based on orthogonal Procrustes alignment (Williams et al., 2021), and Procrustes with isotropic scaling (Schönemann, 1966), measure distance up to rotation and scale; ρ_{sim} is the Procrustes problem with scaling restricted to a k-dimensional projection of the source, and Table 5 reports both. Latent functional maps (Fumero

et al., 2024) and normalized space alignment (Ebadulla et al., 2025) compare and align representation spaces of different dimensionality through spectral or pairwise-distance structure; they answer "how similar are two spaces" whereas the decomposition answers "how much of this space is a projection of that one", and the two are complementary (an NSA-type distance between Z and $Z_{\|}$ would be a natural alternative to our utility metrics). Ye & Lim (2016) give distances between subspaces of different dimensions, which we use in the form of principal angles to compare realized subspaces.

L Limitations in Detail

Global realizability is the wrong criterion for curved data: the true coordinates of a Swiss roll have $\rho_{\mathrm{lin}} = 0.40$. Section I gives the local version, in the spirit of the local dimension estimates of Chun et al. (2026); its neighborhood size and local dimension are choices, and the PCA reference level is needed to read it on anisotropic data. ρ_{lin} is uninformative for $n \lesssim D$ and must be reported against its null; the same applies to relative realizability with a feature source of dimension p: when $p \gtrsim n$ the in-sample residual collapses by Proposition 3 regardless of structure, which is why §F evaluates rich sources out of sample with a regularized map. ρ_{sim} is a nonconvex problem solved to a local optimum, so reported values are upper bounds. The reflection check of Proposition 6(iii) is conclusive only at points fixed by the symmetry; on curved decoded manifolds a local Jacobian spectrum cannot rule out a global discrete symmetry that moves points, and our evidence against mirrored reconstructions rests on the held-out reconstruction error, which Proposition 5 makes decisive. Principal angles between realized subspaces are meaningful only in a data-weighted (score-space) sense on anisotropic data, since unweighted angles are dominated by low-variance directions (Appendix N). What we say about autoencoders (what they identify, when they beat PCA, whether they fix their decoded manifold) is established for MLP autoencoders trained with squared loss, plain, contractive or variational, and should be read as a statement about that class. Our autoencoders are shallow and our network small; our neighborhood embeddings are t-SNE, UMAP and PaCMAP; our natural-image test is CIFAR-10 at 32×32, with pixels as source. Deeper models, pretrained feature sources, and higher-resolution or biological data are the natural next tests, and the released scripts make them routine. Finally, the vocabulary here is a refinement, not a claim that the literature has misused "dimensionality reduction"; the contribution is the decomposition and what it reveals.

M Computation of ρ_{sim}

Let $C = X^\top X$, $B = X^\top Z$. For fixed Q the optimal scale is $s(Q) = \max\{\operatorname{tr}(Q^\top B)/\operatorname{tr}(Q^\top C Q),\, s_{\min}\}$ with $s_{\min} = 10^{-12}$, which enforces $s > 0$ and $f(Q) = \|Z\|_F^2 - 2s\operatorname{tr}(Q^\top B) + s^2\operatorname{tr}(Q^\top C Q)$ has gradient $2(s^2 CQ - sB)$. We iterate $Q \leftarrow \operatorname{polar}(Q - \eta\nabla f)$ with $\eta = 1/(2s^2\lambda_{\max}(C))$, where $\operatorname{polar}(Y) = U_Y V_Y^\top$ from the thin SVD, recomputing s each step, for at most 3000 steps or until the relative decrease is below 10^{-10}, from $Q_0 = \operatorname{polar}(X^+ Z)$. On the analytic case $Z = XU\operatorname{diag}(1,3)$ with isotropic Gaussian X the predicted $\sqrt{2/10} = 0.447$ (at $s = 2$) is recovered as 0.4465 with $s = 1.998$; for $Z = 2.5\,XU$ the solver returns 0. On rank-deficient data (constant MNIST pixels) it leaves residuals of order 10^{-3} for exactly realizable maps, which we treat as zero.

N Experimental Details

Synthetic tier. (a) $X \sim \mathcal{N}(0, I_{10})$, $n = 500$, $W \in \mathbb{R}^{3\times 10}$ Gaussian; U, A from thin QR of $W^\top$. (b) $A = Q_a \operatorname{diag}(\sigma) Q_b^\top$ with Haar-orthogonal Q_a, Q_b and σ log-spaced in $[1, \kappa]$; 20 draws per κ; 2000 random pairs. (c) `make_swiss_roll`, $n = 2000$, noise 0.05, embedded by a Haar-orthogonal 20×3 frame plus $\mathcal{N}(0, 0.05^2)$ noise; t-SNE with PCA initialization and perplexity 30; UMAP defaults; random $\tanh$ features: top two principal components of $\tanh(XW_r)$, $W_r \in \mathbb{R}^{20\times 8}$ Gaussian. (d) $X \sim \mathcal{N}(0, I_{50})$, t-SNE with random initialization and perplexity $\min(30, \lfloor (n-1)/3 \rfloor - 1)$, three seeds per n.

Real-data tier. Pixels in $[0, 1]$; $n = 5000$ images sampled without replacement from each test set (seed 0); inputs centered. PCA by SVD. Random orthonormal projection: Q-factor of a Gaussian $784 \times k$ matrix. Linear autoencoder: `MLPRegressor`, one identity hidden layer of width k,

Adam, learning rate 10^{-3}, batch 128, at most 200 epochs, tolerance 10^{-5}, patience 10. Nonlinear autoencoder: widths $(256, k, 256)$, $\tanh$, Adam, learning rate 10^{-3}, batch 128, at most 300 epochs, tolerance 10^{-6}, patience 20; the bottleneck activations are the representation (a ReLU bottleneck at $k = 2$ collapsed to dead units in a preliminary run). t-SNE: PCA initialization, perplexity 30. UMAP: defaults. Trustworthiness at 10 neighbors; 10-NN accuracy on a stratified 30% held-out split; reconstruction error $\left\|X - \hat{X}\right\|_F / \|X\|_F$ on centered data. *Score-space angles.* On anisotropic data the unweighted principal angles between $\mathrm{span}(X^+Z)$ and the PCA subspace are dominated by low-variance directions that barely affect the code (for the linear AE at $k = 2$ on MNIST the unweighted maximum angle is $41°$ while the codes are indistinguishable from PCA scores); we therefore report principal angles between $\mathrm{col}(Z_\parallel)$ and $\mathrm{col}(XU_{\mathrm{PCA}})$ in $\mathbb{R}^n$, which weight directions by the data variance they carry. *Tangent checks.* With column-vector conventions, $h = \tanh(W_2^\top z + b_2)$ and $\mathcal{D}(z) = W_3^\top h + b_3$ for the stored weights $W_2 \in \mathbb{R}^{k\times 256}$, $W_3 \in \mathbb{R}^{256\times D}$, so the decoder Jacobian is $J_{\mathcal{D}}(z) = W_3^\top \mathrm{diag}(1 - h^2)\, W_2^\top \in \mathbb{R}^{D\times k}$ and the decoded tangent plane is $\mathrm{col}(J_{\mathcal{D}}(z)) \subset \mathbb{R}^D$; it is evaluated at 200 random training points; local data tangent from PCA of the 60 nearest neighbors in pixel space with 20 components; the reported angle is the largest principal angle between the k-dimensional decoder tangent and the 20-dimensional local tangent (0 iff contained).

Reconstruction Jacobians (Table 10). $J = d(\mathcal{D} \circ E)_x \in \mathbb{R}^{D\times D}$ is accumulated analytically by the chain rule through every layer, $J = \prod_\ell W_\ell^\top \mathrm{diag}(\sigma'(a_\ell))$ in forward order, from the pre-activations a_ℓ of a forward pass; $\sigma' = 1 - \tanh^2$ for $\tanh$ and $\sigma' = \mathbf{1}[a > 0]$ for ReLU (the kink at 0 is reached with probability zero). Eigenvalues are computed with `numpy.linalg.eigvals`.

Similarity indices (Table 5). Linear CKA is computed in feature space as $\|Z^\top X\|_F^2 / (\|X^\top X\|_F \|Z^\top Z\|_F)$ on centered matrices. Mean canonical correlations are the mean singular values of $U_X^\top U_Z$, where U_X and U_Z are orthonormal bases of the column spaces obtained by thin SVD; CCA_{50} keeps the 50 leading singular directions of X and $\mathrm{CCA}_{\mathrm{full}}$ keeps every direction with $\sigma_i > 10^{-10}\sigma_1$ (642 for MNIST, 784 for Fashion-MNIST).

Richer sources (Table 8). The images and the t-SNE, UMAP and autoencoder layouts are exactly those of Table 2. Random ReLU features: $\max(XW_r + b_r, 0)$ with $W_r \in \mathbb{R}^{784\times p}$ Gaussian scaled by $1/\sqrt{784}$ and $b_r \sim \mathcal{U}(-1, 1)$, $p \in \{784, 2000\}$. RBF Nyström: `Nystroem` with $p \in \{784, 2000\}$ components and $\gamma = 1/(2 \,\mathrm{median}\, \|x_i\|^2)$. Linear expansion: XW with $W \in \mathbb{R}^{784\times 2000}$ Gaussian, which has the column space of X and 2000 regression features. Supervised features: hidden layer 3 of a $(256, 256, 128)$ ReLU `MLPClassifier` trained for 60 epochs on the first 4000 images. The held-out protocol uses 3500 training points, of which the last 20% serve as the validation split for $\lambda \in \{10^{-3}, \ldots, 10^4\}$, and 1500 evaluation points; the bias term is not penalized; held-out kNN uses the stratified 30% split of every other kNN number, applied to the 1500 evaluation points.

Extensions (§H, Tables 4–3). PaCMAP: `pacmap` 0.9 with default parameters and `random_state=0`. Autoencoders of Tables 4 and 11: JAX, $784 - 256 - k - 256 - 784$, $\tanh$ hidden and bottleneck units (the VAE encoder outputs μ and $\log \sigma^2$ linearly), linear output, Adam with learning rate 10^{-3}, batch 128, 100 epochs, trained on the first 4000 images of the Table 2 subsample and evaluated on the other 1000; the contractive penalty uses the exact encoder Jacobian by reverse-mode differentiation. The decoder-on-PCA control trains only the decoder, on the PCA-k scores of the training split. CIFAR-10: the 10,000 test images, RGB, flattened to $D = 3072$ and scaled to $[0, 1]$ (obtained as JPEG files from a public mirror, so pixel values carry mild compression noise); all methods as in §6.1 except that the autoencoder is the JAX model above trained on all 10,000 images for 100 epochs; trustworthiness is evaluated on a fixed random subset of 5000 images, all other quantities on the full set.

Local realizability (Table 12). 300 anchors drawn without replacement; neighbors by Euclidean distance in the source; local coordinates from `PCA` with $\min(d, K - 2)$ components plus an intercept column; the reported value is the median over anchors.

Depth tier. `MLPClassifier`, hidden widths $(128, 128, 128, 128)$, ReLU, Adam, batch 128, 60 epochs, trained on 4000 of the 5000 images and evaluated on the remaining 1000; hidden activations by forward passes with the learned weights; residuals and participation ratios on all 5000. For the

linearization of Table 14 the affine map is ridge-regularized with λ chosen from $\{0, 10^{-3}, \ldots, 10^{2}\}$ on a 1000-point validation split of the training part, and the bias term is not penalized.